STATE OF WASHINGTON
DEPARTMENT OF ECOLOGY
P.O. Box 47600 • Olympia, Washington 98504-7600
(360) 407-6000 • TDD Only (Hearing Impaired) (360) 407-6006

# Sediment Management Standards (SMS) Rule Revisions
# Freshwater Sediment Standards
# Review Comments and Responses

One of Ecology's rulemaking goals is to adopt chemical and biological criteria for freshwater sediments. Over the last five years, Ecology has been working with other agencies, organizations and individuals to resolve a wide range of technical and policy issues associated with freshwater sediment criteria. Under the draft SMS rule revisions, the biological and chemical criteria would initially be used to establish sediment cleanup standards that are protective of the benthic community. The following materials may be helpful for the reader:

- Background information. We have prepared a short background paper that summarizes the key issues identified during the review process that are associated with the draft chemical and biological criteria. In addition, a high level summary of recent reviewer comments from 2009 through September 2011 is included. This document is titled "Appendix A: Summary of Technical and Policy Issues Associated with Chemical and Biological Criteria for Freshwater Sediments" and is attached to this document.

- Draft rule language. We have prepared a new section (WAC 173-204-573) that includes chemical and biological criteria for the freshwater sediment standards rule language. The updated rule revisions document titled "Preliminary SMS Rule Language" can be accessed at the following website link: http://www.ecy.wa.gov/programs/tcp/regs/2011-SMS/adv-comm/mtg-111028/Handouts/111209-mtg-mat.html

- 2011 SQV Technical Report. The technical report describes the information and methods used to develop the draft chemical criteria (sediment quality values or SQVs). This report titled "Development of Benthic SQVs for Freshwater Sediments in Washington, Oregon, and Idaho" can be accessed at the following website link: http://www.ecy.wa.gov/programs/tcp/regs/2011-SMS/adv-comm/mtg-111028/Handouts/111209-mtg-mat.html

- Peer Review Comments. A collation of reviewer comments to questions Ecology questions posed, individual comments not related to Ecology's questions, and Ecology's responses from 2009 through September 2011. This document is titled "Appendix B: Additional Sediment Workgroup Comments and Responses" and "Appendix C: Additional Peer Review Comments and Responses" and attached to this document.

- Additional Comments. This includes additional comments and Ecology’s responses received during the development of the freshwater sediment standards through September 2011. This document is titled “Appendix D: Additional Comments and Responses” and is attached to this document.

# Appendix A: Summary of Technical and Policy Issues Associated qith Chemical and Biological Criteria for Freshwater Sediments

**Background:**

Currently the SMS rule outlines specific standards and decision-making processes to protect biological resources and clean up contaminated sediment. The SMS rule includes adopted marine numeric chemical and biological standards for protection of the benthic community but lacks adopted freshwater chemical or biological standards. Instead, the SMS has only a narrative standard for freshwater sediment.

There are many contaminated freshwater sediment sites in the state under Ecology or EPA oversight. Because of the lack of adopted freshwater sediment standards, the narrative standard requires a site-specific evaluation to determine cleanup standards. This site-specific process can create inconsistency on how freshwater sediment sites are cleaned up. In addition, the lack of adopted freshwater sediment standards limits how the EPA uses the SMS rule at federal sediment cleanup sites in Washington.

**How have the SMS been revised to add freshwater sediment standards?**

Ecology reviewed a number of options before making a decision on adopting freshwater sediment standards for evaluation and cleanup of contaminated sediment including:

- Develop and adopt only biological criteria.
- Develop and adopt only numeric chemical criteria.
- Develop and adopt both numeric chemical and biological criteria.

After careful consideration, and substantial input from the Sediment Workgroup and MTCA/SMS Advisory Group from 2009 – 2010, Ecology decided to move forward with adoption of both numeric chemical and biological criteria. This decision was based on the need to be consistent with the current SMS framework and to support the most consistent and efficient process for assessing and cleaning up freshwater sediments. Implementing the SMS with only biological or only chemical standards for freshwater sediments would result in higher analytical costs for sediment evaluations and inconsistent evaluations and cleanup decisions.

**What sections of the SMS rule will be revised to incorporate freshwater standards?**

Part V – Sediment Cleanup Standards has been revised to include the freshwater sediment standards for cleanup. WAC 173-204-500 through 590 establishes a cleanup decision process and policies, hazard assessment and site identification, and cleanup standards. Biological tests and endpoints and chemical criteria that establish both sediment quality standards (SQS) and cleanup screening levels (CSL) have been added to Part V, in a new section -573.

In addition, the marine SQS (chemical and biological criteria) in Part III section -315 and -320 have been added to Part V, as a new section -572. This addition does not change the marine chemical or biological SQS or CSL numeric criteria. Instead, it allows the reader to reference Part V for both marine and freshwater SQS and CSL chemical and biological criteria.

**Do the freshwater sediment standards include effects from bioaccumulative chemicals?**

The proposed numeric chemical and biological criteria have been developed to be protective of toxicity to the benthic community. The proposed criteria include some bioaccumulative chemicals but were not developed to be protective of higher trophic levels or human health from bioaccumulative effects. This is consistent with the current SMS framework for marine criteria. However, the revised rule does include rule language for addressing bioaccumulative impacts to human health and the environment. These revisions are focused in new sections -571 (human health protection) and section -574 (aquatic life protection).

**What are the main policy choices Ecology made to develop the freshwater standards?**

- <u>Consistency with the SMS framework.</u> Ecology decided to develop freshwater standards consistent with the current SMS framework, specifically related to:
    - <u>Minor adverse effects</u>. Protection of the benthic community rather than individual aquatic animals or species. This results in allowance of some adverse effects to aquatic animals but overall protectiveness of the benthic community.
    - <u>Two tier criteria:</u> Development of two levels of criteria (SQS and CSL) so a cleanup standard can be established within a range of effects from no adverse effects to minor adverse effects.
    - <u>Applicability to sites:</u> Ecology recognized that freshwater environments are very diverse and that a single set of criteria could not be applied universally at all sites. Ecology decided to develop chemical criteria that were applicable and representative of the majority of sites in the state.
- <u>Biological criteria – use of new sediment toxicity tests</u>. The draft SMS rule revisions include seven sediment toxicity test endpoints to implement the biological criteria. The rule currently provides flexibility for Ecology to approve the use of newer methods as they are validated. Ecology has approved the use of newer marine bioassays on a programmatic level through the annual review meetings, which has been incorporated into guidance, and on a site specific basis. Ecology clarified rule language and has begun outlining supporting guidance to better identify the site manager's discretion in selecting and implementing new bioassay methods.
    - <u>Biological criteria – choice of minimum detectable difference</u>. The draft biological criteria consider both statistical significance and a biological threshold of effects (such as a 10% difference between the test results and the control sediments). There are concerns about the use of biological thresholds greater than 10%. The 2011 SQV report has been revised to provide more detail on this issue. Based on discussions with experts from national labs, it was determined that the initial performance standard for control and interpretive criteria were within the statistical noise for some tests. ASTM round robin tests revealed higher minimum detectable differences that were reflected in setting performance standards and interpretive criteria.

- <u>Chemical criteria – applicability to metals-only sites</u>. The chemical criteria were developed using paired chemical and biological results from a majority of sampling stations that have both metals and organic contaminants. It has long been recognized that the data used to develop the chemical criteria may not be representative of metals-only sites that often have low levels of organic contaminants. While the rule has flexibility to develop site specific criteria and use of the biological override when the chemical criteria may not be representative, there are concerns the language is not clear enough to implement as an ARAR or to set site specific criteria. Rule language was added to provide clarity and specifically identifies metals only sites as an example of when the criteria may not be representative.
- <u>Chemical criteria - protectiveness</u>. There are some concerns about whether the chemical criteria are protective enough. Some of these concerns are related to the policy choice of the use of a 20% false negative rate. The 20% false negative rate for a single test can be interpreted as less conservative. We considered that the 20% false negative rate for a single test does not correspond to the same false negative rate for the combined SQV set or for a single sampled station. The SQS (lower effects level) for each chemical is established by the most sensitive bioassay for that chemical. When combined, the full set of criteria will be lower and result in fewer false negatives than those from individual tests from which they were developed. We also considered that the rule requires multiple bioassays for each station and at least three independent stations to exceed criteria to identify a cleanup site. It was determined that the false negative rate for a single test was more conservatively balanced by these requirements currently built into the rule. These requirements will remain to retain a higher level of conservatism.

**What review did Ecology undertake in the development of freshwater standards?**

Ecology has conducted extensive scientific peer review during and after development of the freshwater sediment standards. The most recent peer review process (2009 – 2011) included the Sediment Workgroup, MTCA/SMS Science Panel, and an external scientific peer review. Ecology considered this input and revised the SQV report in 2011. Below is a summary of the main topics that were commented upon by all reviewers. Appendix B, C and D include more detailed comments.

**<u>MTCA/SMS Science Panel Review</u>.** Ecology asked the MTCA/SMS Science Panel to review the draft 2010 SQV report and accompanying documentation. The panel includes five members: Drs. Bruce Duncan (EPA); Elaine Faustman (University of Washington); Teri Floyd (Floyd/Snider); Michael Riley (Anchor QEA); and Rosalind Schoof (Environ). For more detailed information regarding the meetings, refer to the following link: http://www.ecy.wa.gov/programs/tcp/SciencePanel2009/SciencePanel_hp.html. The panel met in 2010 and 2011 and the main conclusions from these discussions include:

- Science Panel members concluded that the biological tests and endpoints are consistent with the best available science. Like the national peer reviewers, they emphasized the importance of having the flexibility to use newer toxicity tests that are validated in the future.

- Science Panel members agreed that the multivariate statistical approach was a common technique and consistent with best available science. They provided similar cautions on the interpretation of results (associations vs. cause-effect) discussed by the national peer reviewers.
- Science Panel members expressed general support for the approach and analysis. However, they raised some of the same questions/issues about the underlying database identified by the national peer reviewers.
- Science Panel members concluded that Ecology's proposal represents a robust approach and its scientific defensibility is linked to the judgment and flexibility associated with the overall SMS implementation framework (combination of chemical and biological criteria, use of new scientific information, bioassay override of chemical criteria, etc.).

**Sediment Workgroup Review.** From 2009 – 2010, Ecology convened an advisory group for the SMS rule promulgation which included eight local and nationally known experts specializing in sediment management and cleanup. Members of this group included: Clay Patmont, P.E. (Anchor QEA), Glen St. Amant (Muckleshoot Tribe), Dr. Lon Kissinger (EPA), Joanne Snarski (Port of Olympia), Dr. Pete Rude (City of Seattle), Paul Fuglevand, P.E. (Fuglevand, Dalton, Olmsted), Dr. Teresa Michelsen (Avocet), Dr. Jack Word (NewFields, N.W.). The advisory group met three times to discuss the draft 2010 SQV report and the main conclusions from these discussions include (see Appendix B for more detailed comments):

- Chemical criteria – General.
  - Sediment Workgroup members expressed general support for the approach and analysis. As with the current SMS rule and use of the marine chemical and biological criteria, the scientific defensibility of the criteria is based on the flexibility currently built into the SMS rule framework. They raised a number of issues identified above which were worked through, further discussed during Workgroup meetings, and subsequent revisions made to the 2011 SQV report.
  - To retain protectiveness and address freshwater variability, the SMS rule should include both the proposed chemical and biological criteria, the biological override should be retained, and the ability to establish site specific chemical criteria with the use of bioassays should be more clearly established in the rule.

- Chemical criteria – Protectiveness. The work resulted in more reasonable reliability and reduced false positives compared to other SQV sets. This was an important factor for regulatory decisions and for consistency with the current SMS paradigm.

- Chemical criteria – Normalization.
  - **Comment.** Organic Carbon Normalization - Some members expressed the preference to organic carbon normalize polar organics to stay consistent with the marine criteria, EPA guidance, and address bioavailability.
  - **Comment.** BEHP recalculation using OC-normalized values to compare against dry weight values to determine the most reliable method since it is widely distributed and likely to be key chemical for background calculations.

    - **Ecology Response**: Reliability of dry weight and OC-normalized values were compared and dry weight values were consistently more accurate in predicting presence or absence of toxicity.

- Chemical criteria – Summation.
    - **Comment.** Methods for summation of chemicals and how decisions were made to include or exclude chemicals in the FPM was determined to warrant greater detail.
    - **Ecology Response**: Sections were added to the SQV report to describe how reliability testing was used to select from among different methods for treating individual PAHs, total PAHs, TPH.

- <u>Biological criteria – Suite of bioassays.</u>
    - **Comment.** The Sediment Workgroup concurred with the proposed suite of bioassays indicating this represents a best available combination of well developed bioassays and endpoints.  It was recommended to continue to seek additional bioassay organisms and endpoints as they become available.
    - **Ecology Response**: Ecology acknowledged and agreed with the recommendation to move forward with the proposed suite. The rule also includes language that allows flexibility to incorporate new methods based on latest available science subject to Ecology approval. Options that may arise will also be reviewed on a programmatic basis during the annual Sediment Management Annual Review Meetings held each May.

- <u>Biological criteria – statistical and biological significance and reliability</u>.
    - **Comment**. The proposed SMS biological criteria consider both statistical significance and a biological threshold of effects (such as a 10% difference between the test results and the control sediments).  The SQS for acute endpoints for amphipod and midge bioassays exhibited poor reliability when set at 10% difference from control. The 10% difference appeared to have been below the minimum detectable difference (MDD).
    - **Ecology Response**. Ecology consulted with labs and national experts to determine achievable minimum detectable differences (MDD) for bioassay endpoints, especially for the acute mortality tests.  When the SQS effects level was set at the experts' recommended levels, these substantially reduced the variability and when integrated into FPM model runs, resulted in improved reliability of the SQVs. Ecology revised the 2011 SQV report to include a new section that discusses minimum detectable differences, which was one of the factors considered in selection of the biological criteria.

**<u>External Scientific Peer Review</u>**.  Ecology asked four national sediment experts to review the 2010 draft SQV report and accompanying documentation.   The four experts were Drs. Allen Burton (University of Michigan), Jay Fields (NOAA), Chris Ingersoll (USGS) and David Mount (EPA).  Some of the main conclusions and/or observations include the following (see Appendix C for more details):

- Biological criteria – General.
    - **Comment.** Overall, we heard that sediment bioassays are a good tool for predicting sediment toxicity and the biological criteria are technically sound and consistent with approaches being used by other agencies. They are not perfect, but as one peer reviewer stated "it's the best you can do at present".
    - **Ecology Response.** Ecology has proposed rule language that includes freshwater bioassays and retains the biological criteria override of chemical criteria that exists for marine criteria.

- Biological criteria – New tests and endpoints.
    - **Comment.** Reviewers urged Ecology to provide the flexibility to use new bioassays as they become available (freshwater mussels and mayfly tests were specifically mentioned). Three of the four reviewers suggested that Ecology consider using biomass as an endpoint for the *Hyalella* 10-day test, which takes into account both survival and growth.

    - **Ecology Response** Ecology clarified existing language in the SMS rule to emphasize the flexibility to add and modify bioassays as science progresses. We are evaluating the biomass endpoint and will be collecting the data needed to compare with growth and mortality endpoints.

- Biological criteria – Statistical and biological significance.
    - **Comment.** The proposed SMS biological criteria consider both statistical significance and a biological threshold of effects (such as a 10% difference between the test results and the control sediments). There are concerns about the use of biological thresholds greater than 10%. Reviewers agreed that interpretation criteria should consider both statistical significance and a biological threshold of effects, such as a 10% difference between the test results and the control sediments.

    - **Ecology Response** Ecology revised the 2011 SQV report to include a new section that discusses minimum detectable differences, which was one of the factors considered in selection of the biological criteria. Based on discussions with experts from national labs, it was determined that the initial performance standard for control and interpretive criteria were within the statistical noise for some tests.

- Chemical criteria – General.
    - **Comment.** Reviewers agreed that multivariate statistical models provide a credible basis for developing chemical criteria. However, most of the reviewers cautioned that these models represent associations (not cause-effect relationships) between sediment concentrations and biological effects. Reviewers also thought that the data quality requirements were scientifically sound.

    - **Ecology Response** Ecology ensured the revised 2011 SQV report did not have language that inferred a cause and effect relationship.

- Chemical criteria – Database.
    - **Comment.** Ecology asked reviewers if they thought the database provided sufficient data to develop statewide chemical criteria. There were a number of questions on the geographic distribution of the sediment results and the large percentage of results for certain bioassays from Portland Harbor. However, at least one reviewer said he did not expect different concentration-response relationships in different geographic areas.

    - **Ecology Response** Information was added to the 2011 SQV report to show the geographical distribution and hits and no-hits data distribution. While Portland Harbor area did dominate some of the bioassay datasets (up to 75% of the data for *Hyalella* 28-day growth), the new map shows reasonable geographical distribution across Washington for samples and hits.

    - **Comment.** About 20% of the bioassay results represent "hits" (failed the bioassay). Three of the reviewers raised questions about the low percentage of hits and the implications for criteria development. Two reviewers suggested that Ecology examine this issue by recalculating the criteria by using the stations with hits + an equal number of stations with no toxic responses. We could then check the reliability of the resulting chemical criteria using the remaining no-hit stations.

    - **Ecology Response** The floating percentile method used to develop the chemical criteria is not adversely impacted by hit/no-hit ratios. However, reliability evaluations are impacted by the lack of hits in the dataset. Several options were considered on how to deal with this issue, and Ecology adopted suggestions from Burt Shepard of EPA. Multivariate models reflect correlations between variables (in this case, the correlation between sediment toxicity and chemical concentrations). The validity of these types of models is judged by their ability to accurately predict toxicity. EPA suggested that we use three additional reliability measures (Bias, Odds Ratio, and Hanssen-Kuipers Discriminant). The analyses using these reliability measures indicate the floating percentile method is protective and reliably predicts sediment toxicity. This new analyses of reliability was added to the 2011 SQV report which evaluates reliability in a manner unbiased by the hit/no-hit ratio.

- Chemical criteria - Protectiveness.
    - **Comment.** Several reviewers recommended that Ecology carefully consider whether the chemical criteria were protective enough. Some of these concerns related to the choice of a 20% false negative rate. Other concerns seemed to stem from the way the floating percentile method works and the potential for some chemical values to become artificially high due to co-occurrence of several chemicals.

    - **Ecology Response** A case study analysis was conducted using alternatives that included use of other SQVs such as TELs and TECs which are lower and have a very low false negative rate. These alternatives were analyzed regarding how they would perform under the proposed rule, which allows a bioassay over-ride of chemical criteria. Results showed the proposed standards and TEC/TELs resulted in similar areas requiring cleanup, indicating that the proposed standards are as protective as the TEL/TEC approach when a bioassay over-ride is allowed. This is due to the high false positive rate of the TEL/TEC, which is directly associated with the lower false negative rate.

## Appendix B: Additional Sediment Workgroup Comments/Ecology Responses

The following are comments and Ecology responses provided in May and June, 2010 by members of the Sediment Workgroup on their review of the 2010 SQV report and a wrap up of the Sediment Workgroup discussions. For detailed transcripts of the Sediment Workgroup meetings, refer to the following link:http://www.ecy.wa.gov/programs/tcp/regs/2009MTCA/SedMtgGroupInfo/SGMtgInfo/SediWGMeetingInfo.htm

**Pete Rude Comment.** Executive Summary: The fourth goal of the report is "Obtain consensus among the RSET agencies on how the SQG calculations and reliability analysis should be conducted, along with the final values". Can you refresh me on Ecology's ultimate goal on SQGs in light of the Sediment Evaluation Framework (SEF)? Is the intent/hope that the freshwater SQS/CSL analyte list and criteria that end up in the SMS are the same as those in the SEF? (Also, what is the equivalent intent for the biological tests used and related endpoints/criteria?)

**Ecology Response.** It was the intent of the RSET interagency workgroup that the freshwater sediment standards would be consistent to the extent possible across states and programs in the region. The specific analyte list that is chosen for use in each program and the regulatory framework around the SQS/SL1 and CSL/SL2 values may differ among programs depending on their implementing authorities and program needs. For example, the dredging programs focus on the SQS/SL1 for open-water disposal, while the WA SMS cleanup program uses the CSL/SL2 for site listing purposes.

Similarly, it is intended that the biological QA/QC and interpretive criteria for the bioassays tests would be common among the programs. Each program (Ecology SMS, DEQ cleanup, ID cleanup, RSET) will need to follow its own adoption and public review process for any values or interpretive criteria selected; the SQV report is being provided to all of the programs as a supporting technical document.

**Pete Rude Comment.** Table ES-1. Recommended Sediment Quality Guidelines: There are obvious differences between the Table ES-1 analyte list and the marine standards chemical criteria list. The report discusses some of these differences obliquely (e.g., TPAH vs. individual PAHs), but there is no general discussion of the differences and why they exist.

**Ecology Response.** These differences exist for several reasons:

- The marine standards were calculated in 1988, using a much different data set, both chemically and biologically. These data are from different regions of the state with different industries and contaminant sources. Therefore, the contaminants that may be present and the levels at which they may exhibit effects to marine benthos may be different from freshwater environments.

- The marine standards would likely change if they were updated today, as there are new biological tests being used in the programs as well as new classes of chemicals being analyzed in sediments. However, neither the regulated community nor the agencies has seen this as a high priority, since the existing standards seem to be working well.

- The marine standards were calculated using an entirely different mathematical approach than the current freshwater standards. The AET approach was evaluated for use in freshwater, but due to much greater variations in bioavailability of chemicals (especially) in the more widely varying freshwater environments, it was found to be unsuitable due to a high false negative rate.

- Because the FPM is a fully empirical approach, the model results themselves indicate which chemicals are important to include, based on which chemicals contribute to toxicity in the model. For all the reasons noted above, we did not constrain the model runs to the chemicals previously included in the marine standards.

**Pete Rude Comment.** If the FW SQGs are promulgated in some form, what will the ramifications be of having different analyte lists in the two parts of the revised SMS? Some sort of policy statement or discussion might be helpful. The report describes well why there is no TOC normalization, and it seems like a discussion of how the analyte list was developed and what it means in the context of the existing marine chemical criteria is in order.

**Ecology Response.** As noted above, it would be expected that the analyte list and the levels would differ somewhat between the two different environments, although there may be larger variations than expected due to the different mathematical approaches. At Ecology's direction, the above summary could be included in the report or in a policy discussion that accompanies the report or the rule-making effort. The SMS does require the most protective standard be applied where different media come in contact (e.g., groundwater and sediments, or surface water and sediments) and this same requirement would hold true where freshwater sediments and other media, including marine sediments come in contact.

**Pete Rude Comment.** What would be the standard suite of chemical tests to address the SQG analytes? I see at least the following if a full suite is called for: Metals, SVOCs, PCBs, Organotins, TPH (two methods), Conventionals (ammonia, total sulfides, and probably TOC), and Grain size. I'm assuming that the rule (or related guidance) will state that testing would only be done when there is reason to believe the analytes might be present? Is that Ecology's expectation? Will there need to be a special analysis for the pesticides on the list?

**Ecology Response.** As noted above, final analyte lists may differ among programs. For example, ammonia and sulfides were found to contribute to toxicity in the overall data set. However, these chemicals would be lost to the water column during dredged material disposal, and thus there would be no need to include them in that program, as they would no longer be present in sediments at the disposal site. Organotins are already identified as "chemicals of special concern" for dredging projects, and are only analyzed when specific criteria are met regarding uses at the site or existing data.

Each program will make its own decision regarding which chemicals to include on its standard list, and which might be chemicals of concern for only some sites or projects. These decisions would be incorporated into rule or guidance prior to implementation.

**Pete Rude Comment.** Are the SQS levels identified on the table likely to lead to detection limit issues for some of the analytes? If so, it would be good to communicate that and help parties prepare.

**Ecology Response.** As far as we know, all of the SQS levels should be detectable under normal circumstances, as these values were derived from detected historic data for the region. All non-detect values were removed from the dataset prior to the model runs.

**Pete Rude Comment.** Conventional Pollutants (i.e., ammonia and total sulfides) make the recommended list. Are we really ready to have an SQS for these chemicals? What is the typical natural concentration of these chemicals in freshwater sediments? Did RSET do some background work on why these should be included?

**Ecology Response.** These pollutants were indicated by the model to contribute to toxicity in freshwater sediments, which is why they were included in the report. While SQS and CSL values have been calculated for them, it is up to the agencies and each program to determine their use and applicability. It is likely that these standards would have greater applicability to in situ sediments (e.g., cleanup sites) than to dredging projects. In addition, judgment should be applied to determine whether these compounds in the environment are likely anthropogenic based on the site history and conceptual site model, or natural. Both anthropogenic and natural levels can cause toxicity, but only anthropogenic contaminants are subject to cleanup regulations.

**Pete Rude Comment.** Will the FW SQGs rule include the “Nonanthropogenically affected sediment quality criteria” language that is included for the marine standards? It just seems like there is the potential for confusion and extra work/analysis if typical natural levels of ammonia and sulfides are anywhere near the proposed SQS.

**Ecology Response.** The SMS rule does include the “nonanthropogenically affected” subsection, and it applies to both marine and freshwater standards. Rule language revisions are also being considered that would make this section consistent with the MTCA definition of “natural background.”

**Pete Rude Comment.** The addition of TPH is also of potential concern. My understanding is that some natural materials can result in a detection of TPH. Are there enough data available to understand the typical levels of TPH that we might find in un-impacted areas? And, how would the SQS compare to that level?

**Ecology Response.** These levels are well above natural levels of TPH in the environment, and revised TPH protocols that will be issued along with the SMS rule revisions will ensure that natural compounds present in the sample are subject to cleanup procedures prior to analysis.

**Pete Rude Comment.** Section 2.1.5 Bioassay Tests and Endpoints. What freshwater biological tests and endpoints are Ecology actually considering be included in the revised rule? Is the plan to follow the marine regulation – 2-acute/1 chronic – approach?

**Ecology Response.** Please refer to the materials presented to the SMS Advisory Group by Russ McMillan on the proposed biological interpretive framework for the rule revisions. The approach is based on the marine standards, but differs somewhat due to the more limited variety of freshwater bioassays available for use.

**Pete Rude Comment.** Section 2.3 Reliability Analysis. Where did the reliability goals shown in Table 2-3 come from? Has Ecology already approved of them? How do they compare to the goals set for the marine standards?

**Ecology Response.** The reliability goals were set and approved by the RSET interagency SQV workgroup, in which Ecology participated. The goals were based on policy considerations (acceptable levels of error) as well as past performance of the model (realistically possible levels to achieve for SQVs using the FPM approach). Reliability goals were not set for the marine AETs in advance. However, review of the 1988 report indicates that sensitivity ranged from 45-93%, efficiency ranged from 37-100%, and overall reliability ranged from 50-96%, depending on the bioassay. The reliability goals set and achieved for the freshwater bioassays were more stringent than these (80% and above for all measures).

**Jack Word Comment.** Note that comments were provided as redlines within the April 2010 draft of the SQV report. Certain comments indicated that other comments had been answered by later text, and these pairs of comments are not included below. Where necessary, contextual information has been added in brackets.

**Jack Word Comment. General.** Good job! I have a number of comments embedded in the text addressing areas where I have concerns. I am not sure we solve all of them in the near term but want us to keep thinking about how to handle some of these “hidden dragons” that we have yet to confront.

**Ecology Response.** Thank you for your thoughts. The embedded comments are addressed individually below.

**Jack Word Comment.** Page ES-2, second bullet. I am concerned at three levels when SQG are established based on biological response to sediment samples.

- First, there is an assumption that our chemistry suite includes or is a good surrogate for the causative factor(s) that result in the observed toxicity. If the cause(s) of the toxicity are not measured (e.g., pyrethroids) then the suite of measured contaminants are assumed to represent a mixture that caused the observed response. If all stations have the same relative grouping of contaminants (similar source) then this is not a problem. If however, there are multiple sources in the watershed then other chemicals become the surrogate for the actual cause and this may be different between watersheds.

**Ecology Response.** It is one assumption of this and other empirical methods that the measured chemical suite includes the chemicals responsible for at least the substantial majority of the observed toxicity, and it is possible that this assumption is not always correct, i.e., chemicals not measured could be responsible for a portion of the observed toxicity in a manner that does not correlate well with one or more chemicals that are included. Areas with dissimilar sources are purposely included to include a variety of different types of mixtures in the data set and better distinguish among the toxicities associated with the chemicals that are measured. Therefore, if unmeasured chemicals were causing a portion of the toxicity, they would likely contribute to some of the false negatives or false positives that remain once the model has been optimized.

However, it is generally not possible to know with any specificity the cause of the errors that remain. The best that can be done in the meantime is to minimize the errors using the data you have, and if additional chemicals are measured and included in the data set, the model can easily be rerun to see if error rates can be reduced by including them. If in the meantime the station-specific error rates are acceptable, the lack of unknown chemicals can be assumed to not have a major effect on the results, at least within the geographic area represented by the data set.

**Jack Word Comment.** The floating decimal [sic] approach uses the biological responses to classify mixtures of chemicals and their concentrations that provide the best fit of toxic and non-toxic samples by correctly identifying toxicity on the suite of co-occurring bioassay/chemistry determinations. This is a tautological assessment until a new suite of data is compared to an original FPM SQG to see how accurate the prediction was.

**Ecology Response.** Ecology agrees that an independent validation study is needed, but also agrees with the previous RSET workgroup decision to use all the available data to calculate the values, and follow up with a validation study once further data are available. It should be noted that most or all national SQV sets also went through these phases; seldom is it possible to independently validate an SQV set at the same time that it is being calculated.

**Jack Word Comment.** Historical AET and other SQG that have been developed in one area are not good predictors for another area. Since the same test organisms are often used in these different areas with different responses to the chemical specific guideline the lack of concurrence indicates the sediment chemical value is not a good predictor of the biological effects.

**Ecology Response.** The concept behind development and use of regional SQVs has been that different areas may exhibit different responses to the same bioassays. The lack of concurrence in difference areas may reflect a variety of natural factors, such as geochemical differences in the sediments being tested, and testing factors, including different bioassay organism stocks or laboratories. Therefore, it is considered that SQVs are more likely to be accurate if they are developed based on a regional, rather than national, data set, and are used only within that same geographic area. The only other options are other empirical SQVs based on other areas of the country (such as the Great Lakes) or nationally, which are less likely to be accurate.

**Jack Word Comment.** The toxic effects of contaminants are really based on the "bioavailable" fraction. An extensive body of work has indicated that non-polar organic contaminants in the dissolved state are the toxic and bioavailable fraction of the contaminant and this is driven by TOC. Certain metals are controlled by AVS levels, others by association with Al or Fe, others by microbial processes, and in the case of pyrethroids it's whether there is the presence of PBTO to activate the process. Normalization to these attributes of sediment is as important, if not more important than normalizing to dry weight of sediment.

**Ecology Response.** The assumption that the only bioavailable fraction is the dissolved fraction holds better for contaminants moving into the water column or bioaccumulation pathways. Toxicity testing studies with benthic invertebrates have frequently confounded this assumption by showing that feeding pathways and direct exposure to and/or ingestion of sediment may also be important. This may be the reason that evaluations using large field data sets such as this one have never shown an improvement in reliability with any kind of normalization, although it has been tried many times over the years. None of the other available SQV sets are normalized, and to date, no improvement in reliability has been seen for FPM values over 10 years of projects. It should be noted that historical data are not available to attempt certain forms of normalization, such as AVS-SEM, nor is there any data on PBTO with which to assess pyrethroid availability.

**Jack Word Comment.** The final attribute is the quality and quantity of food to sediment dwellers. Not all of them use what is in the sediment but rely on overlying water transport of food materials that are captured out of the water column. Assuming that all sediment dwellers feed on what is in the sediment is not a good assumption. Poor survival, growth or reproduction may be a response to insufficient food in sediment or released from sediment into the water column.

**Ecology Response.** Feeding issues continue to be under evaluation and discussion for several bioassays. There is little that can be done to address this while calculating SQVs from historic data; however, improvements in future bioassay protocols may allow this source of uncertainty to be reduced in the future.

**Jack Word Comment.** Table ES-1, footnote on "greater than" values. Bioassays often do not show effects when SQG are exceeded, so many groups are willing to biologically test sediment. Which should be the controller, biological response with non-contaminant issues accounted for or exceedance of a SQG based on historical information where non-contaminant issues have not been accounted for? Is this allowed within the currently proposed SMS?

**Ecology Response.** Under the framework currently in use for the marine standards and dredging projects, bioassay results would always override chemical results or exceedance of chemical SQVs. The same interpretive framework would be used for the freshwater standards. This option allows either the agency or the regulated party to use more direct measurements of toxicity if they believe that any of the issues in the above comments may affect the reliability of the SQVs. However, because the false positives associated with the FPM SQVs are much lower than with alternative SQV sets, the disconnect between the SQVs and bioassay results should lessen.

**Jack Word Comment**. Section 2.1.2, Completeness**.** In other regions it has been found that pyrethroids are significant contributors to the toxicity of sediments. Since pyrethroids are difficult to measure at effects based levels they are not generally examined in Washington. It was demonstrated in San Francisco and Newport Bay watersheds – even small drainage areas that the pyrethroids were a predominant cause of the toxicity and not the other chemicals normally measured. Special assessments were made using TIE approaches and also examining for the presence of PBTO that is added to improve the efficacy of the pyrethroids. It is likely that the switch from chlorinated to OP pesticides and then to pyrethroid based pesticides has impacted our biological tests without being looked for in sediments. This is a good example of where SQGs for other chemicals are then acting as a surrogate for pyrethroid related toxicity.

**Ecology Response.** Walla Walla District in particular is concerned that sediments in that area may be impacted by OP pesticides. In areas where such chemicals may be anticipated based on surrounding land uses, it would make sense to use modified or additional testing procedures to better assess the potential causes of toxicity. Such site-specific testing is provided for in the SMS based on the conceptual site model. For dredging projects, it would be based on a reason to believe that such chemicals are likely to be present.

**Jack Word Comment.** Section 2.1.2, Minimum Amount of Data. See above comment. Pyrethroids would not be detected at effects based levels, even if looked for so they would be excluded as a potential cause of toxicity, decreasing the reliability of SQGs developed from other chemicals.

**Ecology Response.** If present in the sediments included in the data set, pyrethroids could contribute to the remaining errors once the model has been optimized. However, because the resulting error rates are reasonable, the effect should not be unduly large in most areas.

**Jack Word Comment.** Section 2.1.2, Non-Toxicity. [Fe] Can affect bioavailability of a number of metal contaminants as can Al and Mg.

**Ecology Response.** True, however, chemicals were included as potential toxicants in the model only if they themselves are considered toxic to benthic organisms at concentrations found in the data set.

**Jack Word Comment.** Section 2.1.2, Chemistry Quality Assurance. Somehow we need to determine how to incorporate the allowable level of analytical variation for chemical analyses permitted with acceptable QC. As an examples the % recovery of acceptable spiked compounds into blanks and matrix blanks for organic chemicals is 5-fold (30-150% recovery) and for many metals is 3-fold. Samples within one batch of analyses can be accepted QC-wise with this degree of variation. The same types of variation are accepted with performance standards assessment for certified reference materials. I think this says that a given value in a sediment can have a very significant range, just analytically – we are not even considering within sample variation. We might consider incorporation a spiked or surrogate recovery correction – or at least an assessment or see if that is an issue for odd sample hits or non-hits.

**Ecology Response.** For odd sample hits or an unusual pattern of results, this would be within the latitude of a site or project manager on a site-specific basis. However, for ease of program implementation, a more straight-forward approach to accepting data as accurate (i.e., passing QA requirements) and then using the resulting value directly will likely continue to be used.

**Jack Word Comment.** Section 2.1.3, Normalization and Summing, first paragraph. This needs to be fixed, not the normalization issue. I totally disagree with this viewpoint and it needs to be addressed directly with site-specific measurements. If there is a decrease in predictive ability of organic carbon normalized concentrations of non-polar organic compounds then there is a problem with the guideline that was developed, not the demonstrated relationship that has had extensive research during its development.

**Ecology Response.** The relationship referred to has not been demonstrated with field data. It appears to work at a theoretical and laboratory level, but not with all the variables present in the field. If there is no improvement in reliability after various kinds of normalization, this implies that the theory does not entirely translate to the field, not that the SQVs require correcting. No SQVs derived using field data in North America are organic-carbon normalized, or normalized in any other manner. This suggests that our experience is not unique. Furthermore, this has been tested with a wide variety of different project data sets over the last 10 years, and several times on this project alone.

**Jack Word Comment.** Section 2.1.3, Normalization and Summing, Page 6, fourth bullet. Congener sums are way better predictors of total PCBs by Aroclor – we can discuss this but the method of defining Aroclor mixtures is not trivial and assuming that each Aroclor can be summed without over-accounting for certain congeners in each mixture is not a good assumption.

**Ecology Response.** Unfortunately, we don't have historic congener data to use in deriving SQVs, but agree that this would be a better method of summing if it becomes available in the future.

**Jack Word Comment.** Section 2.1.3, Normalization and Summing, Page 6, seventh bullet**.** Especially for things like dioxins or PCBs the concentrations can be summed after TEF conversion with and without the non-detected values to see what the influence would be. For these organic measures a 5-fold difference in may not be truly different based on the previously discussed QC issue.

**Ecology Response.** TEF conversation does not seem appropriate for assessing toxicity to benthic invertebrates, since the TEFs are based on toxicity to vertebrate species.

**Jack Word Comment.** Section 2.1.5, Bioassay Tests and Endpoints, second paragraph. Need to remember the effect of holding times on the results of Microtox.

**Ecology Response.** Comment noted – this will be addressed in the recently updated Microtox protocol if Microtox is one of the bioassays selected for inclusion in guidance. Microtox was not used to develop the SQVs due to QA problems with the historic data.

**Jack Word Comment.** Section 2.1.6, ANOVA Analyte Screening, last paragraph. Retene and certain degradation products are chemical signatures of coniferous wood burning and not petroleum. They do have toxicity, especially with early life history of fish – through direct water borne and uptake into tissues. How about pyrethroids?

**Ecology Response.** While it may have toxicity through water column pathways, it did not appear that in this data set retene had measureable toxicity to benthic invertebrates. Pyrethroids were not detected and/or not analyzed, and therefore, could not be included in the model.

**Jack Word Comment.** Section 2.2, SQG Calculation - Floating Percentile Method. I like this approach as a standard but would really like to see how SQGs created in rivers of Washington compare to SQGs values created for California, Alaska or Florida. Concurrence among multiple locations would make me way more confident that the SQGs developed here are real. Are there any cross comparisons yet. When this was done for the AET and its application from Puget Sound to San Francisco or San Diego the concurrence was poor, indicating that something else was really causing the observed effects (variation in bioavailability of the contaminants, other contributing factors to toxicity, etc.).

**Ecology Response.** There have been a number of past projects using the FPM, including Oakland Bay, Los Angeles Harbor, Onondaga Lake, and Portland Harbor. However, not all of these data sets were freshwater, and some of them (e.g., Onondaga Lake) were unique, with only a few contaminants. However, it is not assumed that SQVs would be the same from region to region or that that would be a measure of their accuracy – quite the opposite. This approach is designed to reflect regional variations in geochemistry, bioavailability, and bioassay testing to derive region- or site-specific guidelines that would be expected to differ from place to place. The SQVs are also mixture-derived, and thus are influenced in part by the specific industries and other sources present in a region.

**Jack Word Comment**. Section 2.2, Page 13, Step 1**.** How about adding an allowable range or the range that has occurred with QA accepted data within the database? Also what do we do with things like pyrethroids that cause toxicity below the chemical detection limits? Otherwise the process looks good.

**Ecology Response.** The range described here is the range of concentrations that exists among the QA-accepted data in the database, once non-detects and rejected data are removed. At this point, the model has no way of handling chemicals that may be causing toxicity below detectable levels, as there is no way to know whether they are present at all. In addition, few if any of the available surveys analyzed for pyrethroids.

**Jack Word Comment.** Section 2.4, PAHs vs. TPH. The total petroleum hydrocarbon explaining more than the total PAH values seems strange to me. Most of the oil work I have been involved with suggests the opposite – I wonder why?

**Ecology Response.** Theoretically, TPH should encompass greater toxicity than PAHs alone, since PAHs are a subset of the compounds included in the TPH analysis. Aromatics are only one source of toxicity to benthic invertebrates, which are also sensitive to narcotic toxicity caused by aliphatic compounds.

**Jack Word Comment.** Section 2.5, Final Model Runs, second bullet. Good! [referring to older Microtox data being removed due to data quality issues]

**Ecology Response.** Comment noted.

**Jack Word Comment.** Section 2.6, fifth paragraph**.** Really? [referring to there being no ESA-listed species in areas where projects are likely to be conducted in WA, OR, and ID]

**Ecology Response.** Yes – NOAA and USF&W conducted a search for the RSET workgroup to ensure that there were no ESA-listed benthic species in WA, OR, or ID in areas where projects were likely to be conducted. There are ESA-listed freshwater mussels and snails, particularly in ID; however, their range does not overlap with areas in which cleanup or dredging projects have historically occurred or are likely to occur.

**Jack Word Comment.** Table 3-7, middle header (Distribution of Floating Percentile Values). Can you add titles for the columns?

**Ecology Response.** This area of the table presents the distribution of SQVs for each chemical calculated for the various bioassays. The values for the different bioassays are presented from low to high, with between 4 and 10 values per chemical. There are no specific headers for the columns, since each distribution is different and may be in a different order. We will explain this section further in a footnote.

**Jack Word Comment. Section 4, Conclusions, sixth bullet.** The comment was that there were none. I am wondering about freshwater mussels – I believe they are protected and threatened or endangered throughout their range?

**Ecology Response.** See response to Section 2.6 comment above.

**Clay Patmont Comment.** Following up on my preliminary thoughts e-mailed to the Sediment Workgroup on May 2, provided below are my comments on the draft freshwater sediment quality guideline (SQG) report cited above. Overall, I think the SQGs developed in the draft report do a commendable job of developing reasonable reliability (~80%) and controlling false positives, and the document represents a strong step forward to develop freshwater sediment criteria.

**Ecology Response.** Thank you for the comment.

**Clay Patmont Comment.** However, there are at least several issues that should be addressed more thoroughly in the final report, including:

- Organic carbon (OC) normalization;
- Chemical criteria for bis(2-ethylhexyl)phthalate;
- Use of reference samples for confirmatory biological determinations;
- Future incorporation of additional sediment data; and
- Other miscellaneous comments.

Each of these issues is discussed below.

**Ecology Response.** Thank you for your comments; each is addressed below.

**Clay Patmont Comment.** Organic Carbon-Normalization. The draft report recommends use of dry weight-normalized SQGs based on the following:

- Dry weight SQGs have similar or potentially better reliability compared with OC-normalized values;
- OC-normalized SQGs can be difficult to understand and explain to the regulated community; and

- In some situations, anthropogenically derived organic carbon can complicate OC normalization.

**Ecology Response.** It should be noted here that the first bullet above was the primary consideration. The second bullet refers to the difficulty describing and implementing OC-normalization given the high variability in TOC within a site and across the data set. This has been a constant source of uncertainty with the marine standards and does not warrant revisiting for freshwater standards when OC normalization was determined to be less reliable. Had the OC-normalized criteria shown better reliability, the other issues could have been addressed with some effort, as they have been in the marine program.

**Clay Patmont Comment.** However, I submit that a comprehensive weight-of-evidence evaluation would conclude that OC normalization provides a more defensible and consistent set of SQGs for polar organic chemicals including polychlorinated biphenyls (PCBs), polynuclear aromatic hydrocarbons (PAHs), bis(2ethylhexyl)phthalate and others. The weight-of-evidence evaluation would consider the following: The minor differences in reliability reported between dry weight and OC-normalized SQGs do not appear to be statistically significant and thus should not be used as a basis to select between these approaches;

**Ecology Response.** Any form of normalization or other data manipulation would need to provide significantly better reliability than direct use of the data in order to include it in the model. Program experience with OC-normalization has shown that it has problematic aspects that do not warrant its use unless it provides some demonstrable benefit in interpreting real-world data over and above theory.

**Clay Patmont Comment.** The scientific underpinnings supporting use of OC-normalization to assess the potential bioavailability of polar organic chemicals is now firmly established and widely supported in the technical literature and other regulatory programs – e.g., EPA's current sediment quality benchmarks for PAHs and other polar chemicals are based on OC-normalization for this reason;

**Ecology Response.** To our knowledge, OC-normalization is only supported in a theoretical sense and for SQVs developed purely on theory (e.g., equilibrium partitioning values developed by EPA). However, these values are not used in any state regulatory program or by the dredging agencies, all of which use dry weight-normalized values. In addition, equilibrium partitioning theory has not done a good job of predicting toxicity to benthic organisms, likely because they feed directly in or on sediment and ingest particles.

**Clay Patmont Comment.** The most recent EPA guidance on sediment bioavailability assessments for polar organics uses a two- or three-phase partitioning model consisting of the bioavailable dissolved phase and a generally non-bioavailable particulate OC phase (the generally non-bioavailable dissolved organic carbon phase can also be added as a further refinement) – consistency with these other regulatory programs should be maintained;

**Ecology Response.** As noted above, these theoretical approaches are not used in practice by state or dredging agencies. Consistency with the other programs in use in the region would dictate use of dry weight values.

**Clay Patmont Comment.** In addition, Ecology and other state agencies recently collaborated to develop more detailed guidance on how to assess bioavailability of contaminants in sediment – consistency with this guidance can be achieved in part through incorporation of OC-normalization (the web link for this bioavailability report is as follows: http://www.js3design2.com/con_sed_web_jws);

**Ecology Response.** Chapter 4 of this report discusses methods for assessing bioavailability of chemicals to benthic organisms, and does not strongly advise use of OC-normalization. Rather, it is presented as one of many tools that could optionally be used. However, it does emphasize the use of bioassays and benthic community assessments to "infer" bioavailability, which is consistent with Ecology's interpretation framework.

**Clay Patmont Comment.** Within a given sediment site, OC-normalization has been demonstrated to strongly improve predictions of bioavailability for polar organics, but the variability in the types organic carbon between sites (e.g., soot versus wood) contributes to the observed variability in region-wide correlation analyses – the generally accepted bioavailability model for polar organics and a summary of variable sorption characteristics associated with different sediment materials for a representative PAH compound (phenanthrene) is depicted below (note that some forms of anthropogenic carbon such as coal, soot and activated carbon dramatically decrease the bioavailability of polar chemicals, further supporting the use of OC-normalization);

**Ecology Response.** Rather than attempt to predict the bioavailability of chemical compounds using models that have been demonstrated to be problematic in real-world environments, and require the collection and conduct of specialized chemical tests and models, Ecology prefers to use the more direct approach of simply measuring toxicity through bioassays, should either the applicant or the agency suspect the presence of confounding factors that would cause the SQVs to be less predictive.

**Clay Patmont Comment.** The promulgated Sediment Management Standards (SMS) marine chemical criteria for polar organic chemicals are based on OC-normalization - regulatory consistency with the existing marine SMS chemical criteria should be maintained, as there are no clear or compelling reasons not to do so;

**Ecology Response.** While it would be ideal to have consistency between the two, it is also helpful to maintain consistency with the dredging program, which uses dry weight values. Dry weight marine AETs have come into increasing use in the program as problems with OC-normalization have surfaced. Therefore, moving both moving both marine and freshwater standards in the direction of dry weight is more likely than OC-normalization, especially in the absence of any compelling reason to use OC-normalization.

**Clay Patmont Comment.** While occasionally some of the lay public may be confused by OC-normalization, the vast majority of the regulated community readily understands this procedure, particularly since it is the basis for the existing SMS marine criteria, EPA benchmarks, and expanding federal and state guidance; and

**Ecology Response.** On the contrary, while large users such as Port districts may be well aware of the procedure, smaller regulated parties, land-owners, the legal community, and even agency staff are unfamiliar with it. It presents continuing confusion over units, presentation of data, and comparison to other SQVs that are in common use nation-wide.

**Clay Patmont Comment.** Though the presence of anthropogenically derived organic carbon can complicate OC normalization, there are relatively straightforward approaches that are currently applied to address this situation, such as setting lower and upper bounds on organic carbon levels that can be used in the normalization calculations and excluding macroscopic woody materials from the organic carbon analysis.

**Ecology Response.** This is true, however, these procedures are less than satisfactory, as they may result in different stations at the same site being handled differently, and introduce an unnecessary source of controversy and debate that complicates site investigations and interpretation of data, particularly when dry weight interpretations and OC-normalized interpretations differ.

**Clay Patmont Comment.** Thus, OC-normalized SQGs for polar organic chemicals should be used in the final report as the basis for SQS and CSL chemical criteria for polar organic chemicals, similar to the promulgated marine criteria. There should also be a narrative provision included with the rule that further evaluations of bioavailability can be performed as necessary on a case-by-case basis (e.g., using current high-resolution sediment porewater sampling and analysis methods) to provide a more direct assessment of bioavailability.

**Ecology Response.** Based on our program experience and that of other regulatory agencies, as well as the lack of additional demonstrated reliability with this method, Ecology respectfully disagrees with this recommendation. Ecology recognizes that no approach using SQVs will be predictive all the time, and that site-specific factors will play a role in bioavailability. Hence, Ecology prefers to use bioassay testing to assess these factors directly, rather than become overly engaged in predictive evaluations and theoretical discussions that may or may not reflect actual conditions in the field.

**Clay Patmont Comment.** Chemical Criteria for Bis(2-Ethylhexyl)Phthalate (BEHP). The proposed sediment quality standard (SQS) chemical criterion for BEHP (500 µg/kg dry weight basis; this should be recalculated using OC-normalized values as discussed above) deserves additional scrutiny. BEHP is a ubiquitous contaminant in urban stormwater that contributes to regional background conditions, and a contaminant that will likely be a focus of regional and municipal source control efforts in the years and decades to come. Therefore, it is critically important that a reliable sediment quality value for BEHP is selected carefully. The reliability of the current recommended BEHP SQS chemical criterion is suspect given the following observations:

The currently proposed SQS chemical criterion for BEHP (500 µg/kg) is based on the floating percentile model (FPM) for Hyalella 10-day mortality, whereas the FPM for Hyalella 28-day mortality is almost 1,000 times higher (>440,000 µg/kg) – the longer-duration test should at least be as sensitive as the shorter-duration test;

**Ecology Response.** Theoretically, one might agree – if the two data sets were the same and the long-term Hyalella test had been in use as long as the acute test and had as good a track record. However, the difference here may be that there is a much smaller data set for the chronic Hyalella test, which is also distributed in less urban regions of the state. In addition, the variability in the chronic test results is greater, and thus it has not always shown greater sensitivity than the acute test.

**Clay Patmont Comment.** The currently proposed cleanup screening level (CSL) value for Hyalella 10-day mortality (22,000 µg/kg) is substantially higher than the SQS value for Hyalella 10-day mortality (500 µg/kg), suggesting that the test result is overly sensitive to the selection of the adverse effects level;

**Ecology Response.** The data set and the results of a variety of runs were checked to ensure that neither of these values were anomalous or unusual results of the model. Many different runs for the acute Hyalella test were conducted to test different endpoints or approaches to the data; all resulted in SQS values within a factor of 2 of these results, while this was the run with the best reliability of the group (by a fair amount). The data set is robust, with a large number of data points surrounding these values, especially the SQS. These results would suggest that there are effects associated with BEHP at this low level, but that it takes a fairly high concentration to see more moderate effects.

**Clay Patmont Comment.** The proposed SQS chemical criterion for BEHP of 500 µg/kg is likely above [sic, should be “below”] typical reference concentrations in urban areas (note that reference comparisons were not used to develop these SQGs), making this value particularly problematic;

**Ecology Response.** It does not seem likely that a true reference concentration for BEHP would be found in an urban area, since the most common source of this compound is CSO and stormwater outfalls.

**Clay Patmont Comment.** EPA does not list a chronic water quality criterion for BEHP in its National Recommended Water Quality Criteria because “There is a full set of aquatic life toxicity data that show that BEHP is not toxic to aquatic organisms at or below its solubility limit“; and

**Ecology Response.** Water and sediment toxicity are frequently not the same, as sediment-dwelling organisms have different sensitivities to compounds than vertebrates and are exposed directly to sediments as well as to pore water or the water column. It should be noted that if you assume a 1% TOC value, the freshwater SQS is essentially the same concentration as the marine SQS, lending further evidence for the existence of benthic effects at this concentration. In addition, there have been some urban sites near outfalls where bioassays failed and this and/or other phthalates were the only chemicals above the benthic standards (e.g., Duwamish/Diagonal CSO).

**Clay Patmont Comment.** EPA lists the water solubility of BEHP at approximately 340 µg/L - applying the average log Koc for this compound of approximately 5.1 and setting the criterion at 1 percent of the solubility limit to prevent development of non-aqueous phase liquids (similar to petroleum criteria) would result in a calculated solubility-based OC-normalized sediment criterion of roughly 400 mg/kg OC, which is somewhat higher than the current marine CSL chemical criterion of 78 mg/kg OC.

**Ecology Response.** See above response.

**Clay Patmont Comment.** In consideration of the apparent inconsistencies in the sediment quality database, and the lack of demonstrated toxicity in water quality studies, it is likely that BEHP provides little or no contribution to sediment benthic toxicity. Thus, both the SQS and CSL chemical criteria for BEHP should either be eliminated or revised, also folding in the OC-normalization procedure discussed above as appropriate.

**Ecology Response.** Experience with this compound in the marine program does not support no toxicity in sediments to benthic organisms, and the proposed SQG is consistent with that already promulgated for marine sediments. It appears that there were no anomalies in the model calculations for this chemical, and it has a robust data set. Therefore, the proposed SQGs are likely appropriate as they stand.

**Clay Patmont Comment.** Use of Reference Samples for Confirmatory Biological Determinations**.** While control comparisons were used to reduce the noise in the database and to optimize the number of valid data pairs for the purpose of developing chemical screening criteria, it is crucial to clarify both in the final SQG report and in the rule that control comparisons should not be used for confirmatory biological testing evaluations.

**Ecology Response.** Comparison to control will certainly be a default option, as is currently the case for both freshwater and marine sediments, in cases where freshwater reference areas are not available, or where reference comparison is attempted and performance is not acceptable. Many programs across the country use comparison to control as a default, and this option may be the only one available for most smaller sites and dredging projects. For larger sites that have the resources to carry out substantial reference site investigations, comparison to reference may be available.

**Clay Patmont Comment.** On page 7, the possibility of identifying reference sites for freshwater sites is discussed, and confounding factors such as the greater heterogeneity and changing conditions associated with freshwater locations is noted. While this is an accurate summary of freshwater sediment reference characteristics, it also underscores the vital importance of developing appropriate “real world” reference data for confirmatory testing evaluations, since use of controls for such comparisons will result in a significant false positive bias.

**Ecology Response.** This is easier said than done in regions such as ours, with few truly large bodies of water and few uncontaminated upstream areas that share the same geochemical characteristics as the contaminated sites. RSET attempted to locate freshwater reference areas for years, and was not able to do so in any of the three Corps districts. However, RSET did produce a guide for identifying appropriate reference areas providing a step-by-step procedure that could be followed, should candidate areas be identified in the future. In addition, reference sample QA guidelines have been developed as part of the rule development process and are available for use should reference samples be identified and collected.

**Clay Patmont Comment.** While this issue may be beyond the scope of the existing draft SQG report, the final report and the rule should nevertheless clarify that appropriate statistical procedures will be needed to properly interpret confirmatory bioassay data, including use of a reference envelope or pooled reference data approach (e.g., combining suitable reference data across studies, potentially with subsets to address different grain size categories).

**Ecology Response.** This comment will be taken under advisement and should be discussed further at SMARM or other similar venue. Such an approach to reference comparisons would be quite different from that used in the dredging or cleanup programs previously (other than Portland Harbor), but similar approaches may also arise from bioaccumulation work being done programmatically and at various larger sites, leading to development of techniques that could be considered.

**Clay Patmont Comment.** Future Incorporation of Additional Sediment Data. The screening approached used in this evaluation may miss some important studies such as the Ecology Quendall/Baxter data for PAHs or upper Columbia data for metals. These data were apparently screened out based on the observation that "toxicity would frequently occur in samples without adequate chemistry to explain it". In many situations this conclusion may be erroneous (e.g., focused chemistry/bioassay analyses are often based on early screenings which determined that other chemicals of concern were not important). A step should be added that evaluates such cases so that in the future these robust datasets are included in the SQG updates.

**Ecology Response.** Assumptions are often made that limited chemicals are present without adequate data to ensure this, which may be acceptable for site-specific purposes but runs the risk of adding complications and noise to an SQV dataset that is already large and heterogeneous. Because the FPM method depends so heavily on mixture-derived data, it is very important that as many analyses as possible be represented at each station. In addition, the types of sites referred to above frequently have unusual and confounding factors, such as large quantities of wood waste or mining wastes, which are not appropriate samples to include in the data set.

**Clay Patmont Comment.** Other Miscellaneous Comments. The proposed SQS chemical criterion for naturally-occurring metals such as nickel (26 mg/kg) appears to be below the state-wide background (38 mg/kg per Ecology's reports), so this should be adjusted (reference concentrations of other metals should also be checked);

**Ecology Response.** Because the SQVs were developed for a tri-state area, and geochemistry varies greatly even within the state, the RSET committee agreed not to adjust the values in an attempt to account for background, since these concentrations are apparently toxic in some areas. Most of the values are well above background; however, some may be problematic in specific areas. All of cleanup and dredging rules allow for adjustment of standards to background if the natural background is higher than the standard. That said, not all of the metals may be recommended for the final list of SQVs, which is still under discussion and subject to public comment.

**Clay Patmont Comment.** The bulk (i.e., dry weight-based) SQGs for ammonia and sulfide should be removed, as bulk sediment concentrations of these chemicals are only very poorly correlated with the bioavailable dissolved fraction – development of SQGs for these naturally occurring chemicals should be based on porewater concentrations and should also explicitly consider reference conditions (similar to the BEHP discussion above).

**Ecology Response.** As noted above, bioavailability to sediment-dwelling invertebrates includes direct contact with sediments and ingestion of sediments, pathways that may or may not involve dissolution into porewater. Both ammonia and sulfides have exhibited toxicity at pulp mill, wood waste, and other sites in the region, and accounted for some portion of the toxicity observed in the model. However, as has been noted elsewhere, each program may select those contaminants from the list that make sense for their program purposes. Ammonia and sulfides likely would not be included in dredging guidelines, since they would be lost to the water column during dredged material disposal. These compounds are most important for in situ sediments. As with all sediment sites and naturally occurring compounds, judgment would need to be used along with a conceptual site model and site history to determine if it is appropriate to regulate these compounds as contaminants.

**Clay Patmont Comment.** The proposed SQGs for petroleum compounds are also problematic, in part because concentrations similar to the proposed SQS and CSL values are common in organic rich sediments, particularly within urban areas – the following additional steps should be included in any further application of the petroleum criteria, potentially resulting in removal of the criteria:

- Verify that non petroleum organics were removed from the samples through appropriate cleanup steps in the analytical method

**Ecology Response.** Historic analyses of TPH did not always conduct cleanup that would remove all non-contaminant TPH from the samples. However, TPH was nevertheless highly correlated with toxicity in the model, more so than almost any other contaminant. Ecology is issuing new protocols with the standards to ensure that all future analyses of TPH include appropriate cleanup steps. Because the existing SQVs may have included some natural component, they may be higher than they otherwise would have been; however, we do not expect this component to be large in comparison to other fractions present, based in part on an independent analysis of the chromatograms.

**Clay Patmont Comment.**

- Verify with chromatograms or other chemistry review that samples actually contained petroleum rather than just other organics (e.g., wood can frequently give false positives)
- Assess the different types of petroleum and consider parsing the bioassay data out further within the fractions (toxicity can vary greatly with different sub fractions or weathering of petroleum)

**Ecology Response.** An independent analysis of all available chromatograms was conducted by Entrix on behalf of the petroleum industry. This analysis identified two major fractions present – a lower-concentration set of samples dominated by combustion-derived compounds, frequently located in urban environments and near outfalls, and a higher-concentration set of samples dominated by bulk petroleum, located in areas with petroleum sources, such as oil terminals. Because the samples did not significantly overlap in concentration, it would not be possible to calculate criteria independently for the two types of compounds, since a wide concentration range is needed. In addition, the TPH method does not provide the type of data needed to separate out these fractions within a sample without substantial extra analysis by a consultant. Because this would be an added analysis, Ecology and the other agencies are interested in keeping the cost as low as possible and the data simple to interpret.

# Appendix C: Additional External Peer Review Comments/Ecology Responses

## Chemical Criteria

During the scientific peer review of the chemical criteria, Ecology asked that four reviewers focus on the following technical aspects of the technical report titled *“Development of SQVs for Freshwater Sediment in Oregon, Washington, and Idaho”*.

### General Approach:

**Ecology Question 1.** Ecology has developed the draft sediment quality values by using a multivariate statistical model to characterize the relationship between chemical concentrations and sediment bioassay results from sediments collected from freshwater lakes, rivers, and streams in the Pacific Northwest. Do you agree with Ecology’s conclusion that the use of sediment bioassays provides a credible basis for predicting adverse benthic effects that is consistent with current scientific information?

**Peer Reviewer Dr. Burton**: Yes. Well accepted, but in the same context, should only be done in a “weight-of-evidence” approach that utilizes multiple assessment approaches. Please acknowledge up front the limitations of SQGs as noted in Wenning et al 2005 (the SETAC Pellston conference book on SQGs). Among these limitations note the uncertainty associated with assigning causality between single chemical SQG exceedances when the empirical data is from sites with multiple chemical exposures.

**Ecology Response:** The 2011 SQV report was modified to remove any implied causality.

**Peer Reviewer Dr. Fields**: Yes, with limitations. Sediment bioassays currently represent the best available approach for evaluating adverse effects on benthic macroinvertebrates. Their credibility would be enhanced with the development and standardization of other non-lethal endpoints and other test species. Relying on sediment bioassays for predicting adverse benthic effects requires a high quality database of matching chemistry and toxicity and high standards for test quality to minimize false negatives and false positives.

**Peer Reviewer Dr. Ingersoll:** Yes.

**Peer Reviewer Dr. Mount**: All methods of assessing risk to benthos have the potential for either artifacts or misinterpretation, but sediment toxicity tests, used in a synthetic way with other available information, are in my opinion the best available approach.

**Ecology Question 2**. Do you agree with Ecology’s conclusion that multivariate statistical analysis provides a credible basis for characterizing the relationships between chemical concentrations and biological test results?

**Peer Reviewer Dr. Burton**: This statement is too extreme. They assist in the decision making process, but cannot be used alone- particularly given the inherent lack of causality relationships that are associated between the two.

**Ecology Response:** The report was modified to remove implied causality. Additionally, these values are part of a decision making process- the first step in screening. Bioassay over-ride of standards is also a part of the decision process, as is consideration of bioaccumulatives (different section of the rule) in both aquatic /aquatic-associated organisms and with respect to human health where consumption of aquatic organisms may occur.

**Peer Reviewer Dr. Fields**: Yes, empirical models have a long track record in characterizing concentration-response relationships. However, the strength and reliability of these models is increased with evaluation of data independent from the development database.

**Peer Reviewer Dr. Ingersoll:** Yes.

**Peer Reviewer Dr. Mount**: I believe FPM can be a useful tool for gaining insight into relationships between sediment toxicity and sediment chemistry. However, I think the report reflects a misplaced belief that FP analysis indicates which sediment contaminants are "responsible" for sediment toxicity, as opposed to being correlates of toxicity which may be caused by other contaminants. Many statements are made in the report at either imply, or directly assert that certain contaminants are, or are not, responsible for observed toxicity. However, there is nothing at all in this report or the underlying data that speak to causality. Further, some other sources of information argue against such relationships. For example, even though certain metals are found by FP and many other empirical SQG approaches to be a strong correlate of sediment toxicity, TIE studies conducted in our laboratory have never indicated metals to be a cause of toxicity in sediments, with the exception of extreme contamination from very specific sources of metals, such as mining wastes. This is supported by analysis using approaches like (SEM-AVS)/foc or interstitial water analysis, which typically shows that generalized metal contamination common in harbor sediments is not at levels sufficient to attribute toxicity to those metals based on our best understanding of metal bioavailability and toxicology. This does not diminish the fact that there is a strong empirical relationship, but I think it is egregiously wrong to assert that there is a causal relationship based on empirical analysis, and/or that a regulatory approach should treat metals as the "problem" in such sediments, as opposed to (appropriately) recognizing that that level of metal contamination is often associated with sediment toxicity. As further evidence of this issue, the report contains the indication that maybe the proposed standards should not apply to mining sites. The basis for this was not explained, but I infer that it is because the proposed standards perform poorly for data from mining sites. Isn't this a strong indication that these SQVs are not reflecting causality, if they perform poorly at locations where it is almost certain that metals are the cause of toxicity?

**Ecology Response.** The report was modified to remove implied causality. Also note that the basis for not applying the standards to mining sites was not based on poor performance, but a lack of substantial data from mining sites that met the minimum dataset requirements for inclusion in the model. Without a strong dataset, Ecology preferred to incorporate the cautionary note.

**Ecology Question 3.** The data used to develop the draft chemical criteria is summarized in Section 2.1 of the technical report. Do you agree with Ecology's conclusion that the database provides sufficient data to support the development of statewide chemical criteria? That is, is the dataset sufficiently robust for the development of chemical values specific to:

- Geographical coverage.
- Coverage of different types of freshwater systems.
- Numbers of paired chemistry and bioassay endpoints.
- Number of bioassay species.
- Number of acute and chronic tests (referring to test duration relative to life history).
- Number of lethal and sublethal effects endpoints.
- Percentage of hits and no-hits for each endpoint.

**Peer Reviewer Dr. Burton**: The database is limited and needs to be augmented, from many perspectives including: geographically, chemically, concentration ranges, and biological effects (bioassay and indigenous). If one looks at the limited concentration ranges for some chemicals it is obvious that the resulting frequency distributions are weak from a scientific and ecosystem relevance perspective.

**Ecology Response.** The ranges may be limited compared to across the US, but are typical of this region. It should also be noted again that these criteria are used in screening- bioassay over rides are part of the process.

**Peer Reviewer Dr. Fields**: The report did not provide a detailed breakdown by endpoint for the geographic areas or the different types of systems, although I would not expect different concentration-response relationships in different geographical areas or different types of systems. However, it appears that a large percentage of the data for *Hyalella* 28d survival and *Chironomus* survival/growth came from the Portland Harbor Superfund site. Assuming the chemical gradients in Portland Harbor are consistent with other areas, this may not be an issue. It could be a problem if most of the data for specific chemicals came from one area (TPH?). For Chironomus, sufficient data are available to compare chemical distributions from Portland Harbor with the distributions in the remainder of the database.

**Ecology Response.** The types and ranges of contaminants found in Portland are similar to what is found in other industrial sites in the Pacific Northwest. That is, the Portland Harbor site includes a wide variety and distribution of sources that have resulted in sediment contamination where different chemical families (e.g., SVOLs, TPH or metals) occur by themselves in some locations and overlap considerably at others. Even for PAHs, the Portland Harbor sources are very diverse and provide as broad a spectrum of resulting PAH signatures as would be encountered throughout the remainder of the study area. While Portland did represent the majority of data for TPHs, 23% of the synoptic data used in the model came from other systems. For these reasons the Portland Harbor data is thought to be representative of a broader spectrum of sites geographically and chemically.

**Peer Reviewer Dr. Fields**: A table showing the range and number of toxicity values that were classified as non-toxic and toxic for each endpoint and screening level would be useful.

**Ecology Response.** Table 2-3 of the 2011 SQV report includes this data.

**Peer Reviewer Dr. Fields**: The database for *Hyalella* 28d growth is very limited and probably not sufficient to develop models that can be applied with confidence to new data. The rationale for excluding the data for *Hyalella* growth/biomass from Portland Harbor was not clearly explained, especially considering EPA used these data in the Superfund ecological risk assessment and feasibility study. The lack of TPH data in the *Hyalella* 28d growth database was used as a reason to ignore the PAH models for that endpoint in the selection of the final SL1 and SL2 values.

**Ecology Response.** The same *Hyalella* 28-day growth data was also excluded in the Superfund study. PAHs were not excluded in the final screening value selections - they were included as total PAHs (sum of the standard 17 PAHs). Individual PAHs were not included, since they did not correlate well with toxicity.

**Peer Reviewer Dr. Fields**: Given the lack of *Hyalella* 10d growth and limited data on 28d growth, *Chironomus* 10d growth was the only non-lethal endpoint with sufficient data. The growth endpoint was treated as independent from the survival endpoint, although that is not likely to be the case. The biomass endpoint, which takes survival into account, may be a more useful way to assess sublethal effects.

**Ecology Response.** Noted, although insufficient data is available at this time to add this endpoint into the rule, the rule is sufficiently flexible to allow the incorporation of this endpoint into assessments. In the RSET and SMS programs, changes to program endpoints can be made, but follow a specific process. An issue paper is prepared along with supporting data and other evidence, and presented at the public Sediment Management Annual Review Meeting. The federal and state sediment management agencies jointly consider the recommendation and make a decision after the meeting, considering public comment. Subsequent actions, such as rule revisions or database reprogramming, could take significantly longer. This approach would also be used for changes to the biomass endpoint for revision of the bioassay control criteria.
The development of chemical thresholds is limited by the low percentage of hits for all endpoints. The impact of the low prevalence of toxicity on the reliability measures was not discussed in the report.

**Ecology Response.** A new section on reliability was added to the 2011 SQV report, based on Burt Shepard's (U.S. EPA) comments to the same effect.

**Peer Reviewer Dr. Ingersoll:** No. The sources of data used to develop the FPM are not adequately identified. Importantly, the SQVs are to be applied across regions within all three states, not for WA state alone. Insufficient analyses are presented in the report to determine if the proposed SQVs are sufficiently reliable to be used on a regional basis within a state or across the three states.

**Peer Reviewer Dr. Ingersoll:** Geographical coverage. No, insufficient data or insufficient analyses presented in the report.

**Ecology Response.** Figure 2-1 and the revised Appendix A have added geographical information.

**Peer Reviewer Dr. Ingersoll:** Coverage of different types of freshwater systems. No, insufficient data or insufficient analyses presented in the report.

**Ecology Response.** The data was derived from 20 years of cleanup work performed in Washington and Oregon and is representative of the systems most commonly affected by freshwater sediment contamination. These sites are not evenly distributed across the variety of different freshwater systems (e.g., large and small lakes or streams from east or west of the cascades) but they met the targeted goal to increase the breadth of systems represented. Most importantly, the majority of known and suspected sources to freshwater sediment contamination are primarily located on the types of freshwater systems that are best represented by these data.

**Peer Reviewer Dr. Ingersoll:** Numbers of paired chemistry and bioassay endpoints. Possibly, but it seems like most of the data may be from Portland Harbor, which is not a typical site based on ongoing evaluations that are being conducted by USEPA associated with a baseline ecological risk assessment at Portland Harbor

**Ecology Response.** The *Hyalella* 28-day growth is dominated by Portland Harbor, but other endpoints are much more evenly distributed across other regions. The variety of sources and resulting sediment contamination from the Portland Harbor area are diverse and span the normal range of sediment conditions expected to be encountered elsewhere around the region.

**Peer Reviewer Dr. Ingersoll:** Number of bioassay species. Yes.

**Peer Reviewer Dr. Ingersoll:** Number of acute and chronic tests (referring to test duration relative to life history). Uncertain. Microtox should not be included in the analyses. Additionally, the 20-day midge should not be encouraged until ongoing research to refine the 20-day midge method starting with 1st instar larvae has been completed.

**Ecology Response.** Neither Microtox nor the 20 day Midge test was included due to lack of data. However, where this assay has been used in the Pacific Northwest Ecology believes the assay is useful given the lack of freshwater chronic bioassays. After consulting with regional labs and experts who have performed the bioassay, Ecology has concluded the test has achieved appropriate reliability to include in our suite. As with all the bioassays, performance of positive and negative controls will be closely monitored for consistency between studies and labs, and Ecology will continue to work with labs to fine tune protocols for this and other tests.

**Peer Reviewer Dr. Ingersoll:** Number of lethal and sublethal effects endpoints. No. Amphipod 10-day growth should be added and total biomass of amphipods and total biomass of midge should be added.

**Ecology Response.** Data collected from the standard test will allow addition of biomass as an endpoint to the bioassay, and 10-day growth can also be added. However, until sufficient data is available, this endpoint cannot be incorporated into the rule. Flexibility in the rule language will allow addition of bioassays and endpoints when sufficient data is available to allow development of biological assay standards.

In the RSET and SMS programs, changes to program endpoints can be made, but follow a specific process. An issue paper is prepared along with supporting data and other evidence, and presented at the public Sediment Management Annual Review Meeting. The federal and state sediment management agencies jointly consider the recommendation and make a decision after the meeting, considering public comment. Subsequent actions, such as rule revisions or database reprogramming, could take significantly longer. This approach would also be used for changes to the biomass endpoint (see below) or revision of the bioassay control criteria.

**Peer Reviewer Dr. Ingersoll:** Percentage of hits and no-hits for each endpoint. No, insufficient data or insufficient analyses have been presented in the report. It appears there are a low percentage of toxic samples, which compromises reliability estimates. Sample number must be added to all of the estimates of reliability in Tables 3-3 and 3-4.

**Ecology Response.** Ecology adopted Burt Shepard's (U.S. EPA) recommendations for applying statistical evaluations that are independent of the hit/no hit ratio. These new approaches are included in the revised approach and 2011 SQV report.

**Peer Reviewer Dr. Mount**: This is a policy determination, not a scientific one. There is no standard for sufficiency, and no matter how deep the underlying data are, there is a potential for bias and artifact to be introduced by unrecognized quirks in the underlying data. One thing I was a little surprised was not addressed in the report was the degree to which different subsets within the overall data set represented different suites of contaminants. Particularly given that different data sets have different numbers of samples, this seems potentially important.

**Ecology Response.** Each study had to meet a minimum standard set of analytes. They may have more analytes, but not less than the minimum. This is discussed in detail in Section 2.2 of the 2011 SQV report.

**Peer Reviewer Dr. Mount**: With regard to geographical coverage, I don't know of any evidence to suggest that sensitivity of sediment toxicity tests to contaminants has a geographical underpinning. The factors that influence response, such as organic carbon and AVS, can be found across great ranges within most any geographical area I'm aware of.

**Ecology Response.** Comment noted. A number of factors that can affect availability of contaminants such as water hardness, pH, or nature and geological origins of sediments (e.g., igneous or sedimentary formations) are notably different between water bodies east or west of the Cascade range. Due to these influences on contaminants and the potential for different mixes of contaminants from agricultural and industrial activities unique to eastern or western portions of the state, a primary goal in updating the data set was to increase representation of east-side locations.

**Data Issues:**

**Ecology Question 4.** Do you agree with Ecology's conclusion that data was screened using criteria acceptable for the purposes of chemical criteria development? Specifically, refer to section 2.1.2 of the technical report which includes completeness of sediment chemical analysis, minimum number of detected analytes, QA/QC of sediment chemistry and bioassay data, and elimination of chemicals that were not directly associated with toxicity.

**Peer Reviewer Dr. Burton**: Some of the decisions seem arbitrary and contrary to typical, national approaches. There needs to be increased transparency as to why these decisions were made. Why would you use 5 replicates vs. 4. Is WDoE's scientific vetting process superior to ASTM or USEPA's? All of the screening criteria need a strong scientific basis – else you risk the perception that data were deleted to make it work for WDoE.

**Ecology Response.** ASTM recommends between 4 and 8 replicates, and underscores 4 replicates as the "…absolute minimum…" and recommends 8 replicates for routine tests. Moving to higher than the minimum replicates was considered desirable since Ecology applies an overlay of statistical difference. Accepting a minimum of 5 replicates was deemed reasonable minimal increased costs as this was what most labs were doing in the Pacific Northwest. Very few datasets were eliminated by increasing to 5 replicates (only five 10-day Hyalella tests were removed), whereas moving to even greater number of replicates severely limited the number of datasets. Appendix B of the 2011 SQV report describes in detail why datasets were removed and should eliminate concerns regarding deletion of data.

**Peer Reviewer Dr. Fields**: In general, data screening was appropriate and apparently systematic. Other than vague references to an "insufficient analyte list", the criteria for completeness of the chemical analysis were not clearly explained. Similarly, the rules for calculating sums were not explained in the document, particularly whether a specific number of PAHs were required to calculate a sum.

**Ecology Response.** The 2011 SQV report, Appendix B contains details of screening and summing.

**Peer Reviewer Dr. Ingersoll:** For the primary bioassay endpoints, ASTM considers 4 replicates acceptable, so it is not clear why they were excluded. If variance is an issue, that could be addressed. Also, it was not clear that non-toxic samples were evaluated for sufficient power to detect a significant difference. Because of the importance that individual non-toxic samples may have in the FPM approach, such an evaluation should be considered.

**Ecology Response.** See response above to Dr. Burton's comment.

**Peer Reviewer Dr. Mount**: While I have no specific reason to believe it was not appropriate, the rules for screening the data were described in such general terms I don't know that it's possible to say for sure. I will say that I think it is a mistake to exclude data from sediment tests with only 4 replicates, especially without backing this decision with any quantitative analysis. Most of our laboratory tests have 4 replicates and we get detectable differences well within the bounds of the WDOE minimum effect levels. How much data is excluded because of this requirement?

**Ecology Response.** See response above to Dr. Burton's comment.

**Reliability Testing of the Chemical Criteria**:

**Ecology Question 5.** The state and federal agency representatives involved in this effort decided that it was preferred to use all available data to generate the criteria rather than to hold several datasets out of the process for validating the model. A validation study dependent upon new paired sediment chemistry and bioassay data would likely require years of data collection. Thus, reliability was evaluated with the dataset used to generate the values, looking at several reliability endpoints (refer to Section 2.3 Reliability Analysis in the technical report).

Do you believe that the approach used to evaluate the reliability of the criteria is consistent with current scientific methods and principles for validating criteria?

**Peer Reviewer Dr. Burton**: I am particularly concerned that sediment chemistry and bioassay sediment sampling were not co-located given the massive spatial heterogeneity that exists in most sediment sites. Were samples split for chemistry and bioassays? If not – state it. Please do not ignore this huge uncertainty level.

**Ecology Response.** All data were based on the co-located samples which were analyzed for chemistry and bioassays. This was (and is) the first point discussed in Section 2.2 of the 2011 SQV report, initial data screening.

**Peer Reviewer Dr. Fields**: The reliability metrics used are commonly used values, but have serious limitations given the low prevalence of toxicity for all endpoints and screening levels. Other approaches should be included. According to Shepard (2010), endpoints with low prevalence of toxicity (as observed in the freshwater database used) can be "made to meet reliability goals (overall reliability, predicted no-hit reliability and false positive rate) merely by raising sediment quality benchmarks so high that most or all toxic stations are incorrectly classified as nontoxic." Shepard recommends the use of other metrics such as the Kappa statistic that either take the prevalence of toxicity into account or is independent of prevalence.

**Ecology Response.** The revised 2011 SQV report has a new section using Burt Shepard's (U.S. EPA) suggested metrics on reliability.

**Peer Reviewer Dr. Fields**: The reliability measures for individual chemicals should be presented for the selected models and screening levels. Although the individual chemical models are applied as a group, there are likely to be cases where a single chemical model is driving the classification. It would also be useful to show the number of cases for each chemical where that chemical model alone was responsible for the sample classification.

**Ecology Response.** The FPM values are designed to be used as a group, not as individual chemical screening values. This will be made clear in accompanying guidance. For cases where a single chemical is the driver, and there is reason to believe that the standards are not performing adequately as a screen, the rule is flexible enough t allow for over-riding the numerical standard with bioassay testing.

**Peer Reviewer Dr. Ingersoll:** No, insufficient data or insufficient analyses presented in the report. Reliability of individual SQVs needs to be reported.

**Ecology Response.** The revised 2011 SQV report has a new section using Burt Shepard's (U.S. EPA) suggested metrics on reliability.

**Peer Reviewer Dr. Mount**: The measures that are used are relevant and have precedent in the literature. However, they represent only one form of assessment. Another important issue is the degree to which they represent true causal thresholds, since their use is predicated on an assumption that they are directly linked to the presence or absence of toxicity. Because the derivation of these numbers does not address bioavailability, there is a large, inherent error in the values, because one or two dwt based numbers simply can't account for the range in effect thresholds that exist among sediments with different characteristics relating to bioavailability. This is discussed at greater length in my general comments.

**Ecology Response.** Ecology's rule is based on application of screening values, but allows for over-ride of numerical standards with bioassay testing. This flexibility can be used when bioavailability becomes an issue. For example, where there is extremely high TOC or abnormal TOC sources.

**Ecology Question 6.** What comments do you have on the completeness and relative weight that should be given to the various reliability measures used to assess the results of the model and to compare it to other SQV sets?

**Peer Reviewer Dr. Burton**: Why are the criteria not compared to ERM/ERLs or PELs/TELs, etc.? These criteria have been documented to be useful and of similar accuracy to yours, so their comparison is absolutely essential. If you do not do it, someone else will and then report it in the peer-reviewed literature.

**Ecology Response.** This analysis is in the revised 2011 SQV report (Tables 4-2a through e).

**Peer Reviewer Dr. Fields**: The completeness and relative weight of reliability measures depend on the narrative intent of the SQG set. If the goal is to be protective of benthic effects, then a low false negative rate and high predicted no-hit reliability may be more important than false positive rate or predicted hit reliability.

**Ecology Response.** The SMS were not designed to require proof of absence of toxicity. Rather the SMS were developed to manage environmental risk by minimizing adverse effects to a population or community of benthic organisms – the standards are protective of adverse effects in terms of preserving the overall integrity of the benthic community. WAC 173-204-315, “no adverse effects” for marine bioassay tests are defined using effects thresholds of 15-30% compared to reference, depending on the test endpoint. These thresholds are based on the minimum detectable difference in these bioassay endpoints. The freshwater SQVs were designed with the same principles in mind, except they are compared to the control, a more stringent standard. Therefore, 15-20% difference from control is well within the regulatory definition of “no adverse effects” for bioassay tests and is based on minimum detectable difference, or the point at which an adverse effect can first be observed. Effects below this level are considered “no adverse effects,” while effects exceeding this level are considered “minor adverse effects” up to the higher CLS/SL2. As a result, the primary goal stated above is met, as there are not expected to be biologically observable or statistically meaningful adverse effects below this level.

**Peer Reviewer Dr. Ingersoll:** SQVs that are intended to serve as no observed effects levels (SL1) should not have a 20% false negative rate, this rate should be below about 10%. In addition, SQVs intended to represent minor effect threshold should be based on level 1 rather than level 2 effects.

**Ecology Response.** See response to Dr. Fields comment above. Additionally, Ecology would not consider a threshold of 15-20% moderate to severe effects. This would be inconsistent with the current SMS rule, Ecology programmatic practice, and data suggesting that effects in this range represent the minimum detectable difference in most laboratory bioassays. As noted above, the marine biological SQS in Ecology’s SMS rule are based on 15-30% thresholds of effects, representing minimum detectable differences. These levels are also used by the DMMP/RSET regional dredging programs. Washington state has evaluated these thresholds extensively over the years, conducting power analyses and other evaluations to ensure that the thresholds could be detected. This was not the case for all bioassays, suggesting that levels lower than these are not effectively implementable. It is likely that any program that is using statistical significance alone as its hit/no-hit criterion is achieving roughly these thresholds in practice.

While some SQVs use risk-based thresholds rather than effects thresholds due to statutory or mathematical differences from the SMS, thresholds higher than 10% are typically used in both cases. For example, British Columbia uses an EC20 (20% effects level) for development of its sediment quality guidelines. EPA’s Great Lakes National Program Office guidance recommended a difference from control of 20% (http://www.epa.gov/glnpo/arcs/EPA-905-B94-002, Chapter 6). Oregon’s risk assessment guidance sets an Environmental Baseline Value at the LC50 and requires that there be no more than a 10% chance that 20% of the population exceeds the LC50 (which would require a multi-station evaluation). Benchmark values involving the TELs and TECs are often set at 0.2 (20% probability or incidence of a hit). While these approaches aren’t entirely comparable to one another mathematically, we are not aware of any regulatory programs using a 10% or lower threshold on a programmatic (non-site-specific) basis.

**Peer Reviewer Dr. Mount**: I am concerned about the relationship between the number of toxic samples in the database and the performance criteria. For example, the number of toxic samples in the database is about 20%; if the acceptable false negative rate is 20%, then it doesn’t matter where the SQV lies, the false negative rate is likely to meet a requirement of 20% false positives. If the number of toxic samples in the database was 50%, it would exert quite a different pressure on the algorithm. This could be evaluated by redeveloping the SQVs using a database containing all of the toxic samples, but a randomly selected subset of non-toxic samples such that the total number of non-toxic samples was comparable to the number of toxic samples.

**Ecology Response.** Rather than re-calculating values multiple times, Ecology opted to run reliability statistics that were independent of the hit prevalence. See Section 4.3 in the 2011 SQV report for updates to the reliability analyses.

**Peer Reviewer Dr. Mount**: The categorization of samples as simply toxic or not toxic throws away a large amount of information regarding magnitude of effect. This is important – incorrectly classifying a "skinny hit" as non-toxic is a much different error than classifying an egregiously toxic sample as non-toxic. This could be assessed by developing frequency histograms or cumulative frequency distributions for the magnitude of response observed in each of the four classifications within the predicted v observed contingency table.

**Ecology Response.** While bioassay outcomes used in the FPM are represented only as hits or no-hits, the magnitude of toxicity is actually carried into the FPM analyses when you consider how interpretive criteria were established for the bioassays. The bioassay SQS and CSL endpoints are established at the lowest end of the range of observed effects. SQS is set as the minimum detectable difference (MDD) for each bioassay and CSL is established as the lowest discernable increase above that MDD. For example, if the MDD is 15% difference from control, that is where SQS is set, and if the bioassay protocol can statistically discern a 10% difference above SQS, which is where CSL is set. The resulting numeric criteria are protective of the lower end of the observed toxicity range, and satisfy the risk management functions required for regulatory purposes. Once a site has been listed as an area of concern, then Ecology site managers can further address magnitude of chemical and toxicity exceedances using appropriate methodologies including those mentioned in this comment.

**Ecology Question 7.** Are there appropriate alternate validation methods that can use the data from which the standards were developed (i.e., bootstrapping methods, etc.)?

**Peer Reviewer Dr. Burton**: [no response]

**Peer Reviewer Dr. Fields**: The best validation approach would be to apply the values to an independent data set. Without an independent data set, it would be possible to evaluate the impact of the low prevalence of toxicity on the reliability measures by adjusting the relative proportion of toxic and non-toxic samples in the database and deriving new models. For example, assuming 80/20 non-toxic to toxic distribution, it would be possible to randomly divide the non-toxic samples into 4 batches to combine with the toxic samples for comparative model development and reliability.

**Ecology Response.** While Ecology agrees that a true validation with a new, independent, representative dataset would be advantageous, such a dataset is not available and funding is too limited to conduct such a validation. Ecology notes that the SMS marine standards were adopted without a true validation with an independent dataset. Most SQVs are originally developed without validation, which occurs later once new data are available with which to conduct the evaluation. The reliability evaluation is not intended as a replacement for validation, but rather the best that can be done in the meantime. The reliability evaluation conducted for these SQVs is substantial and rigorous and has resulted in values more reliable than other SQVs. The RSET workgroup (including Oregon, Washington, and Idaho state and federal representatives) discussed whether to set aside a portion of the data set for validation or whether to include it all in the SQV calculations. It was decided to include all the available data in the SQV calculations, and to subsequently validate the SQVs once new regional data became available. Ecology still intends to carry out this process.

**Peer Reviewer Dr. Ingersoll:** No, include analyses considering: (1) organic carbon normalization, (2) magnitude of chemical exceedances, (3) magnitude of toxicity, and (4) the influence of different percent difference from control on the percent toxic and on reliability.

**Ecology Response.** (1) new discussion of why TOC was not used is found in section 2.3 of the 2011 SQV report; (2) and (3) the FPM model is not designed to include magnitude of exceedances, chemically or biologically; (4) At the suggestion of the Sediment Workgroup, the impact of difference percent from control was examined for *Hyalella* 10-day mortality endpoint, for which reliability goals were not met. The end result was a change in the hit definition for SQS/SL1 values, which improved reliability.

**Peer Reviewer Dr. Mount**: Why not apply these SQVs to one of the large databases that extend beyond Washington State? I know of no toxicological reason why contaminant response would be regionally influenced. Alternatively, one could add the Washington data to the larger database (probably already in there) then partition the data so that there is a training data set and validation set.

**Ecology Response.** Goals for developing these SQVs were to establish the best predictors of toxicity for mixes and concentration ranges of chemicals encountered in this region. The best test for measuring success in meeting that goal is to examine the reliability or accuracy of these against the regional data. It would be informative to apply to the wider databases but not one of our specific objectives. Additionally, the regional data underwent rigorous QA before being used in the FPM, and this level of review was not available for the national SQV data sets.

**Peer Reviewer Dr. Mount**: In that regard, the comparison of the FP SQVs to other published SQGs seems slanted to me. First, if one really believes that there is a “regional influence” to sediment toxicity, then one should redevelop the other empirical SQGs using only the Washington data to determine whether the FP approach is better. Moreover, the FP SQVs are derived explicitly to combine the various contaminants in a way complimentary to the predictive ability goals. Using other empirical SQGs in a “one hit = predicted toxic” scheme, as was apparently done, is not, in my opinion, consistent with the ways in which the developers of those other empirical SQGs would recommend evaluating the “reliability” of those SQGs for predicting sediment toxicity. My sense is that approaches like SQG “quotients” are the way those multiple SQG assessments are going. To be more fair, it seems to me that the “other” SQGs should be evaluated by developing an SQG quotient (or other combined metric as proposed by the developers of those methods), then using statistical methods to optimize quotients for SQV1 and SQV2 that best meet the predictive goals for those values. That would be apples to apples comparison of the approaches.

**Ecology Response.** The concept behind development and use of regional SQVs has been that different areas may exhibit different responses to the same bioassays. The lack of concurrence in difference areas may reflect a variety of natural factors, such as geochemical differences in the sediments being tested, and testing factors, including different bioassay organism stocks or laboratories. Therefore, it is considered that SQVs are more likely to be accurate if they are developed based on a regional, rather than national, data set and are used only within that same geographic area. The only other options are other empirical SQVs based on other areas of the country (such as the Great Lakes) or nationally, which are less likely to be accurate.

When Ecology first embarked on development of freshwater SQVs, each of the existing SQV sets was evaluated to determine how well they predicted toxicity in Washington and Oregon sediments. The results showed very poor reliability in predicting the toxicity of freshwater sediments in the Pacific Northwest region (SAIC and Avocet 2002). This evaluation was repeated in 2008, with the same results (Avocet 2010). This evaluation is what prompted the agencies to consider development of an alternative approach to setting SQVs that would be more predictive. As was done for the SMS marine standards, Ecology ensured regional sediment samples were collected and analyzed using both Ecology and ASTM approved bioassays. Our goal all along has been to improve on the reliability of existing SQV sets. If at any time we had found that this was not the case, the development effort would have been discontinued in favor of the existing SQVs.

The stated goals for developing the FPM SQVs included the need to improve reliability or accuracy in predicting toxicity for contaminated sediments in our region. The tables examining reliability of different SQV sets and compared their relative ability to predict toxicity in our regional database. Methods like developing quotients to better predict toxicity from a set of SQVs may be useful in evaluating risk at a site, but these were beyond our scope and are very difficult to employ as regulatory standards. The comparisons were performed to best answer the needs for our regulatory program and should not be perceived as an evaluation of the effectiveness of the individual SQV sets in meeting the specific objectives each was developed for.

**Ecology Question 8.** Comparative reliability analysis was used to assess different ways to handle data with respect to number of issues. For example, the toxicity of petroleum compounds can be assessed using various approaches (TPH, individual PAHs, total PAHs, normalizations, etc.), and there were other similar decisions that needed to be made regarding the underlying data set (e.g., summing, normalization, inclusion of conventionals, splitting or lumping geographic areas). Were the reliability comparisons the most appropriate method to make these decisions or are there better methods that could be used?

**Peer Reviewer Dr. Burton**: I have a real problem with WDoE suggesting that their way is better than other national/international ways that have been better vetted through the scientific review process. TPHs have not been shown to be useful – as they are all over the place. WDoE is discounting AVS, TOC, grain size, Fe/Mn/Al as being useful. While none of these are useful at all sites, they certainly can be useful to help with the risk/hazard decision making process at many sites (see the peer reviewed literature). This is disconcerting and will do nothing but raise legal challenges and concerns.

**Ecology Response.** When Ecology and RSET agencies made the decision to develop freshwater standards, it was with the intent to meet the specific regulatory needs of the region. We agreed to use the best available science, updated QA II regional data that reflected the mixtures of contaminants found in the region, and remain as consistent with programmatic practice and the current SMS rule framework as possible. This approach would best answer the needs for our regulatory dredging and cleanup programs and should not be perceived as an evaluation of the effectiveness of other SQVs as each SQV data set was developed with different objectives. After careful analysis and consideration of other SQVs, Ecology determined the FPM approach was an appropriate fit for our regulatory dredging and cleanup framework.

Bioavailability will always differ between sediment samples due to the origin and nature of the chemicals themselves and the many different physical and chemical aspects of the sediment and porewater. A considerable effort has been focused on better understanding the contribution of some of these different factors, such as OC. Ultimately, if there were clear evidence that these other ways of looking at the chemistry served to consistently improve the predictive accuracy of the SQVs, then we would adopt a different approach. The FPM is the only method for developing SQVs that explicitly deals with the bioavailability of a chemical in a synoptic data set by managing its effect on false negatives and false positives. Also, our regulatory framework with a range between the lower and upper regulatory levels and a biological override allows appropriate consideration of site-specific factors where there is reason to believe the SQVs may be less effective predictors of toxicity.

Details on comparisons for TOC, TPH, total PAHs, and individual PAHs have been added in Appendix D of the 2011 SQV report. TOC normalization has been done in the past and in the current calculation effort. TOC-normalization was tested side-by-side with dry weight values in 2003 and in 2008, as well as for a number of other projects, listed in Avocet (2010). These evaluations did not improve the reliability of the results. The marine AETs are the only state or provincial values we are aware of that are TOC-normalized, and it was recognized at the time they were developed that the dry weight AETs were equally predictive. TOC-normalization of the marine AETs was largely done in deference to theory, but it has not been demonstrated to improve predictiveness. In addition, there have been significant implementation difficulties associated with the OC-normalized values, including sediments that had TOC outside the equilibrium partitioning range (too low or too high), sediments with excessive TOC of anthropogenic origin, and general difficulties explaining the approach to the regulated public.

A number of other summing and normalization approaches were also tested, with the result that some chemical classes were summed. Other approaches did not improve reliability. Detailed results from these analyses are now included in Appendix D of the 2011 SQV report.

Lastly, note that these values were designed to function as risk management screening values. Once a site has been listed as an area of concern, either by failing CSL/SL2s or failing bioassays, then Ecology site managers can conduct risk assessment approaches which can include relationships mentioned in this comment.

**Peer Reviewer Dr. Fields**: Reliability comparisons are useful but they only go so far. The underlying datasets for the different approaches differ, so the comparisons are not straightforward. The reliability comparisons were not presented directly, so it was not clear how these were carried out. It would make sense to focus the comparisons on predicted hit reliability and percent of false positives, rather than overall reliability. The only reliability comparisons discussed in the report were between TPH and total PAH, but these were only discussed in a general way. Individual PAHs and organic carbon normalization were apparently dismissed without any comparative analysis.

**Ecology Response.** See above response.

**Peer Reviewer Dr. Ingersoll:** No, insufficient data or insufficient analyses presented in the report.

**Peer Reviewer Dr. Mount**: There are many issues here. Summing PAH concentrations makes toxicological sense, but the document did not describe what would be a minimum set of PAHs that has to be measured for a “total PAH” number to be defined, nor what would be done if a larger set of PAHs (e.g., alkyl PAHs) were measured. Further, it’s not clear that there’s any mechanism for distinguishing between PAH sites that have high or low degrees of alkylation, but it is definitively known that this affects the potency of PAH mixtures. Organic carbon normalization was discarded, which flies in the face of a large body of literature, and I don’t believe the effect of organic carbon normalization was ever examined with the current data set. Even if it doesn’t improve predictive ability, that is not an indication that it “doesn’t matter”, what it means is that there are enough other factors that influence response (e.g., black carbon and other influences on bioavailability) that prevent the organic carbon effect from being recognized by the analysis. Hence, I think it should still be done. I find the “it’s too complicated for regulators” argument to be pretty thin. I absolutely would not separate geographical areas; if the SQV values were regionally dependent, it would be an indication of an unrecognized bias in the underlying data, since there is nothing “regional” about toxicology.

**Ecology Response.** See above response.

## Data Interpretation and Use for Regulatory Decision-Making.

**Ecology Question 9.** Greater differences in the chemical criteria occurred between the bioassays than between SQS and CSL level effects for any one bioassay endpoint. To select SQS and CSL levels for each chemical, the values for all bioassay endpoints and effects levels were combined into a single distribution representing the range of criteria from the lowest no adverse effects level to the highest minor adverse effects level. From this, the SQS was established as the lowest value and the CSL was selected as the next highest, significantly different value (see Table 3-7 of the technical report). This approach was selected as it provides conservative values by remaining at the low end of the no adverse effects to minor adverse effects distribution.

Do you agree with Ecology's conclusion that this approach is consistent with that of the WA SMS marine standards where the SQS and CSL were established as the lowest and 2nd lowest of the Apparent Effects Levels determined for a suite of bioassays?

**Peer Reviewer Dr. Burton**: Seems reasonable from a management perspective. The description of the Floating Percentile and balancing of Type 1 and 2 errors is excessive and in fact confusing. It's not that complicated. The approach is fine and the accuracy rates reported are similar to the other more commonly used SQGs.

**Peer Reviewer Dr. Fields**: I agree that it is consistent in spirit, but there is a much more limited suite of bioassays available for the freshwater database: essentially only 2 species and 3 endpoints-Hyalella survival, Chironomus survival and growth with sufficient data.

**Peer Reviewer Dr. Ingersoll:** I do not have a good enough understanding of the WA SMS marine standards to comment on this question. However, SQVs derived in the report do not represent conservative no-effect levels or minor-effect levels. My experience in evaluating data from other sites indicates that adverse effects would be commonly observed at concentrations of chemicals of potential concern at or below the SQVs.

**Peer Reviewer Dr. Mount**: I have no opinion about this as it would take considerable additional analysis to make such an evaluation.

**Ecology Response.** Due to the greater difference between bioassay tests than between SQS and CSL level effects for one test, a conservative method was used to set the chemical SQS and CSL. All the values for each test endpoint and effects level were combined into a single distribution and the lowest and next significantly different value were selected as SQS and CSL. This approach was determined to best represent the environmentally conservative methods used for the marine standards which had the benefit of a greater number of test species. It is noted that the SQS level and CSL level are established at levels that protect diversity and functions of a benthic invertebrate community. There are more sensitive indicators of toxicity which are most appropriate for assessing risk but are less useful for risk management in the regulatory arena.

**Ecology Question 10.** Given the types of historic sediment data available, is the TPH method the best available approach for assessing the overall effects of petroleum hydrocarbons? What is your answer based on (theory or empirical data)?

**Peer Reviewer Dr. Burton**: No. I have a real problem with WDoE suggesting that their way is better than other national/international ways that have been better vetted through the scientific review process. TPHs have not been shown to be useful – as they are all over the place. WDoE is discounting AVS, TOC, grain size, Fe/Mn/Al as being useful. While none of these are useful at all sites, they certainly can be useful to help with the risk/hazard decision making process at many sites (see the peer reviewed literature). This is disconcerting and will do nothing but raise legal challenges and concerns.

**Ecology Response.** Bioavailability will always differ between sediment samples due to the origin and nature of the chemicals themselves and the many different physical and chemical aspects of the sediment and porewater. A considerable effort has been focused on better understanding the contribution of some of these different factors, such as OC. Ultimately, if there were clear evidence that these other ways of looking at the chemistry served to consistently improve the predictive accuracy of the SQVs, then we would adopt a different approach. The FPM is the only method for developing SQVs that explicitly deals with the bioavailability of a chemical in a synoptic data set by managing its effect on false negatives and false positives. Also, our regulatory framework with a range between the lower and upper regulatory levels and a biological override allows appropriate consideration of site-specific factors where there is reason to believe the SQVs may be less effective predictors.

Details on comparisons for TOC, TPH, total PAHs, and individual PAHs have been added in Appendix D. TOC normalization has been done in the past and in the current calculation effort. TOC-normalization was tested side-by-side with dry weight values in 2003 and in 2008, as well as for a number of other projects, listed in Avocet (2010). These evaluations did not improve the reliability of the results. The marine AETs are the only state or provincial values we are aware of that are TOC-normalized, and it was recognized at the time they were developed that the dry weight AETs were equally predictive. TOC-normalization of the marine AETs was largely done in deference to theory, but it has not been demonstrated to improve predictiveness. In addition, there have been significant implementation difficulties associated with the OC-normalized values, including sediments that had TOC outside the equilibrium partitioning range (too low or too high), sediments with excessive TOC of anthropogenic origin, and general difficulties explaining the approach to the regulated public.

A number of other summing and normalization approaches were also tested, with the result that some chemical classes were summed. Other approaches did not improve reliability. Detailed results from these analyses are now included in Appendix D of the 2011 SQV report.

Lastly, note that these values were designed to function as risk management screening values. Once a site is has been listed as an area of concern-either by failing CSL/SL2s or failing bioassays, then Ecology site managers can conduct risk assessment approaches which can include relationships mentioned in this comment.

**Peer Reviewer Dr. Fields**: TPH method—is the analytical method sufficiently well-defined and consistent throughout the dataset? I am not aware of any data that suggests that same TPH (or total PAH) concentrations from different sources have the same toxicity. Unless a mixture approach is used that can assess the cumulative toxicity of the individual components (e.g., PAH toxic units), treating individual analytes as separate estimators may be the best approach.

**Ecology Response.** All data used the NWTPH methodology. Appendix D of the 2011 SQV report contains details regarding reliability of other approaches (TPH, PAH, individual PAHs.) Note that the ability to sum PAH toxic units would be based on either BaP/TCDD equivalents, which is not a sensitive mechanism in benthic invertebrates, or based on normalizing for narcotic effect, for which sufficient data is not available (would need to normalize to molecular weight of all hydrocarbons, but individual or even grouped hydrocarbon data for full set of narcotic hydrocarbons are not available- TPH was the closest dataset that included the largest group of hydrocarbons). This was agreed to during a multiagency PAH discussion (PAH summit, 6/6/2007; participants included NOAA, USFWS, ODEQ, Ecology, NOAA, USACE, and consultants) - see response to Dr. Mount's and Ingersoll's comments below.

**Peer Reviewer Dr. Ingersoll:** No, insufficient data or insufficient analyses presented in the report.

**Ecology Response.** Section 2.3 of the 2011 SQV report describes the decisions of the PAH summit meeting and Appendix D of the 2011 SQV report details the different methods (Total PAH, TPH-diesel, and TPH-residual) that were evaluated and compared for addressing PAHs. Reliability testing determined the use of TPH always improved predictive accuracy and that the combined use of Total PAH and TPH provided the best results. (Also see the following response to Dr Mount's comment.)

**Peer Reviewer Dr. Mount**: I'm not clear on what is meant by the "TPH method." We have demonstrated experimentally that aliphatic hydrocarbons can induce toxicity to Hyalella and Chironomus separate from the effects of PAHs, so it is very appropriate to include some expression of non-PAH petroleum hydrocarbons among the constituents for which SQV are developed. However, it is not yet clear what the best measure or expression of TPH is with regard to predicting toxicity. I am a little concerned about how the FPM would treat TPH when diesel range and residual are separate variables, as the SQV for one might be raised when it is compensated for by the other. Was there a model run when the two were summed?

Our experimental data indicate that biomass of both *Hyalella* and *Chironomus* are reduced by about 40% at 500 mg/kg aliphatic hydrocarbons (in the form of mineral oil). Based on that, the SQV1 of 340 mg/kg TPH-diesel seems close to right, maybe a little high, while the TPH-residual of more than an order of magnitude higher seems too high, though as I said we don't yet know exactly how best to define the toxicological potency of aliphatic hydrocarbons across a range of molecular weights**.**

**Ecology Response.** Addition of TPH was in part due to a multiagency PAH discussion (PAH summit , 6/6/2007; participants included NOAA, USFWS, ODEQ, Ecology, NOAA, USACE, and consultants), where among other things, the limitations of the standard 17 PAHs were discussed. PAH toxic units approaches were discussed, but the major impact for benthic toxicity is due to narcotic effect, and normalization of all hydrocarbons, not just PAHs, would be needed. Data was not available for this approach. Data for TPH was available for 324 samples (73 of which were outside of the Portland area), and all data were generated using a standardized method used in the Pacific Northwest. TPHs appeared to be better correlated with toxicity within Portland Harbor, and runs with the empirical FPM model (appendix D1) indicated TPHs were better predictors of toxicity in this dataset. Model runs were made using both and reliability testing determined the use of TPH always improved predictive accuracy and that the combined use of Total PAH and TPH provided the best results.

**Ecology Question 11.** Are the measures introduced in the model to assess covariance, coupled with other available statistical tests of covariance, sufficient to address the inevitable co-occurrence of chemicals in the field when developing chemical criteria?

**Peer Reviewer Dr. Burton**: Well, yes and no. Each site is different and that should not be forgotten when making site-specific decisions.

**Peer Reviewer Dr. Fields**: The FPM, like all other empirical approaches, is assessing the toxicity of environmental chemical mixtures where covariance is assumed. The FPM helps to identify the best estimators of toxicity in the database, but it should not be assumed that the suite of values represents a unique solution.

**Peer Reviewer Dr. Ingersoll:** No analyses presented in the report regarding statistical analysis of covariance. Insufficient information presented in the report to evaluate the influence of co-occurrence in sediment chemistry within the database.

**Ecology Response.** A new section on covariance was added to the 2011 SQV report.

**Peer Reviewer Dr. Mount**: Virtually all contaminants show some level of co-occurrence, and I am concerned that the algorithm used may allow SQVs for some contaminants to become artificially high (i.e., exceed a true causal threshold) because other contaminants "cover" for them. This is of particular concern for contaminants whose toxic threshold can be expected to vary widely as a function of bioavailability, such as PAHs and metals. As an example, there are many examples of sediments with PAH-induced toxicity at concentrations well below the PAH SQVs derived here. Yet there are also many examples of sediments with low toxicity of PAHs at very high concentrations because of low bioavailability. Under the FP method, it is possible for the PAH SQV to be allowed to be higher than a true causal threshold because the model wants to eliminate false positives, and the co-occurrence of other chemicals in the "lower but still toxic" PAH samples allows SQVs for other parameters to "cover" the occurrence of toxicity at lower concentrations. This is why it is so important to cross check these empirical SQVs against the known toxicity of these same chemicals to determine how well they agree, and when they don't, to decide what is causing the discrepancy and how best to address it.

**Ecology Response.** A new reliability section was added to the report, which addresses concerns about missing toxicity. Additionally, the rule flexibility allows for bioassay over-ride of the chemical standards if any party has concerns over the standards either missing toxicity, or over-predicting toxicity due to site-specific conditions.

**Ecology Question 12.** Should any chemical classes be summed that were not summed in the model to reduce covariance (e.g. phthalates)?

**Peer Reviewer Dr. Burton**: Only if there is an adequate database (such as for the 16 PAHs) to justify.

**Peer Reviewer Dr. Fields**: Summing chemicals is appropriate when the composition of chemicals included in the sum is consistent, and, when applied, it is necessary that all chemicals included in the sum are measured.

**Peer Reviewer Dr. Ingersoll:** Yes. Metals should be summed as a group (see Ingersoll et al. 2001). Other chemical mixture models should be evaluated.

**Peer Reviewer Dr. Mount**: No strong opinions.

**Ecology Response.** The Ingersoll et al., 2001 method for summing metals is based on normalization to PECs to develop quotients for individual analytes which were then added together. Methods like developing quotients to better predict toxicity from a set of SQVs may be useful in risk assessment at a site, but these were beyond our scope and are very difficult to employ as regulatory standards. Normalizing to an existing screening value, summing, then creating a new screening value did not seem like a logical approach.

## <u>BIOLOGICAL CRITERIA</u>

While reviewing the biological criteria, Ecology asked that the four peer reviewers consider the technical and scientific aspects of using bioassays including bioassay organisms and endpoints. The suite of bioassay species and endpoints were selected based partly on regional availability and familiarity with these organisms.

**Ecology Question 13.** Is the proposed bioassay suite appropriately sensitive to protect the freshwater macro benthic community (i.e., typical taxonomic structure and functions such as a prey base to endangered species like salmon)?

**Peer Reviewer Dr. Burton**: It’s the best you can do at present. Consider adding snails and mussels in the future as they are important and sensitive. Hopefully these document and related policy will be reviewed every couple of years as the science advances.

**Peer Reviewer Dr. Fields**: Are one species of amphipod and one species of midge sensitive representatives of the freshwater macro benthic community (communities)? These bioassays represent endpoints that are currently well-standardized and have matching chemistry and toxicity data available, but no information was presented to assess whether the endpoints used are “appropriately sensitive” or representative of the freshwater macro benthic community. It would be preferable to have additional taxa and sublethal endpoints represented.

**Peer Reviewer Dr. Ingersoll**: Yes, but see comment 53E. Whole-sediment mussel testing should be considered as this method becomes standardized (ongoing research at our laboratory). (Dr. Ingersoll’s comment 53 is inserted here. “ Microtox should not be included in the analyses. Additionally, the 20-day midge should not be encouraged until ongoing research to refine the 20-day midge method starting with 1st instar larvae has been completed.”)

**Peer Reviewer Dr. Mount**: I don't know that there is a great deal of hard evidence that can be brought to bear on this question, but I think it is an accepted presumption that this is true.

**Ecology Response.** Flexibility in the rule allows bioassays to be used in site specific evaluations and standardization is not necessarily a requirement for use on a site-specific basis. Regarding amphipods and midges representing macro-benthic communities, it is recognized that mollusks are not represented but data is not available within the region. Amphipods represent a fully aquatic life cycle and midges, an aquatic larval stage with a terrestrial adult phase, and both represent critical elements in the aquatic food chain. In the RSET and SMS programs, changes to program endpoints can be made, but follow a specific process. An issue paper is prepared along with supporting data and other evidence, and presented at the public Sediment Management Annual Review Meeting. The federal and state sediment management agencies jointly consider the recommendation and make a decision after the meeting, considering public comment. Subsequent actions, such as rule revisions or database reprogramming, could take significantly longer.

**Ecology Question 14.** From your experience, are there other freshwater bioassays/species that provide consistent, reproducible and sensitive results that should be considered for developing biological criteria?

**Peer Reviewer Dr. Burton**: It's the best you can do at present. Consider adding snails and mussels in the future as they are important and sensitive. Hopefully these document and related policy will be reviewed every couple of years as the science advances.

**Peer Reviewer Dr. Fields**: I am not aware of other species that have the same level of method standardization and testing , but USGS is developing freshwater mussel tests that should prove useful.

**Peer Reviewer Dr. Ingersoll**: Mayfly testing (ASTM E1706) and Mussel testing (E2455).

**Peer Reviewer Dr. Mount**: None that have the depth of experience and interpretation behind them. An exception is that I don't see why weight is not included as an endpoint for 10-d *Hyalella* tests (or better yet, combined with survival to calculate a 10-day biomass endpoint). This is not to say that additional tests might not be added in the future as additional work is completed.

**Ecology Response.** The Mayfly test has been suggested as an addition to the proposed tests, but have limitations since there is limited regional experience running the test, they are field-collected and seasonal, so are not being considered for incorporation in the rule language. The freshwater mussel sediment bioassay has yet to be standardized. However, Ecology notes that flexibility in the rule allows additional bioassays to be used in site specific evaluations, and standardization is not necessarily a requirement for use on a case-by-case basis. The reporting of weight for any of the tests can be requested and the results used to look at the biomass endpoint on a site-specific basis. This additional endpoint is of great interest to Ecology and later will be evaluated to determine if it provides an improvement to the proposed endpoints.

In the RSET and SMS programs, changes or additions to program endpoints can be made, but follow a specific process. An issue paper is prepared along with supporting data and other evidence, and presented at the public Sediment Management Annual Review Meeting. The federal and state sediment management agencies jointly consider the recommendation and make a decision after the meeting, considering public comment. Subsequent actions, such as rule revisions or database reprogramming, could take significantly longer.

**Ecology Question 15.** Are there issues you may be familiar with regarding running and interpreting these bioassays (e.g., problems associated with culturing animals, confounding variables that may warrant protocol modifications, choice of toxicant for positive controls, etc.)?

**Peer Reviewer Dr. Burton**: Two primary concerns mentioned under 1 and 2 above (5 vs. 4 reps, sediment chem./bioassay sample splits)

**Ecology Response.** See response above.

**Peer Reviewer Dr. Fields**: No information

**Peer Reviewer Dr. Ingersoll**: Yes, as addressed in ASTM and in EPA standards. Insufficient time to discuss these topics in the context of the report and the timeframe for providing review comments.

**Peer Reviewer Dr. Mount**: There is ongoing work in regard to the appropriate foods and feeding rate for 28-d *Hyalella* tests; there is reason to believe that the standard feeding rate is insufficient to allow maximum growth (Mount et al. unpublished data; some of these data were presented by Ingersoll et al. at the 2010 SETAC meeting). Although we don't yet have a definitive alternative to recommend, we believe we are close to having one and will circulate that as soon as we have it. I'm not aware of a positive control for sediment testing that has any wide use or acceptance.

**Ecology Response.** Flexibility in the rule language will allow addition of bioassays and endpoints when sufficient data is available to allow development of biological assay standards.
In the RSET and SMS programs, changes to program endpoints can be made, but follow a specific process. An issue paper is prepared along with supporting data and other evidence, and presented at the public Sediment Management Annual Review Meeting. The federal and state sediment management agencies jointly consider the recommendation and make a decision after the meeting, considering public comment. Subsequent actions, such as rule revisions or database reprogramming, could take significantly longer

**Ecology Question 16.** What comments do you have on the appropriateness of the various growth endpoints under consideration nationally (e.g., dry weight, ash-free dry weight, length)?

**Peer Reviewer Dr. Burton**: Always do the least variable growth measure to better detect site differences (higher discriminatory power).

**Peer Reviewer Dr. Fields**: My understanding is that ash-free dry weight is required for *Chironomus* (but not *Hyalella*) and either dry weight or length is an appropriate measure for *Hyalella*. Both growth endpoints should be converted to measures of total biomass, which combines mortality and growth into one endpoint.

**Peer Reviewer Dr. Ingersoll**: Biomass of Amphipods and biomass of Midge should be added as endpoints. AFDW of Midge, but not Amphipods should be required.

**Peer Reviewer Dr. Mount**: Rather than assess growth and mortality separately for a given sediment toxicity test, I would strongly recommend using a biomass approach (survival*weight of survivors) to reduce each toxicity test to a single endpoint. The statements made in the report about length being more sensitive than weight are absolutely incorrect. This incorrect assertion is based on a smaller CV for the length measurement. However, the issue is the absolute SD relative to the magnitude of the difference between control and test treatments, and this is small for length. Weight increases with the cube of length, so even though the CV for weight is larger, the dynamic range of the measure is also much larger, such that the power is not different. I can expand on this argument with examples (and a re-analysis of data to show that Stevens et al. 1993 were incorrect in their assertions about length) if needed. No preference should be expressed for length over weight. Weight is also much less time-intensive (and therefore cheaper) to measure than length. <u>Only</u> AFDW should be accepted for *Chironomus*. Dry weight is subject to considerable bias from gut contents. Dwt or AFDW is acceptable for Hyalella. I'm not aware of any evidence to suggest that the choice makes a difference in terms of detecting differences in weight, although given that the dwts are already low, doing AFDW may just introduce additional noise to the measurement.

**Ecology Response.** Language in the report was modified to remove reference to length as an endpoint. Ash-free weights are being evaluated for marine bioassays, and can be done for Chironomids as well. Results from marine systems indicate that this is most needed when sandy sediments may cause increased dry weight in mass. Data collected from the standard analysis will allow addition of biomass as an endpoint to the bioassays; however, until sufficient data is available, this endpoint can't be incorporated into the biological standards. However, flexibility in the rule language will allow addition of bioassays and endpoints when sufficient data is available to allow development of biological assay standards.

**Ecology Question 17:** Is there additional information on the minimum detectable difference for these tests that would assist in setting the SQS (SL1) interpretive endpoint?

**Peer Reviewer Dr. Burton**: None that I know of.

**Peer Reviewer Dr. Fields**: Is there sufficient data in the freshwater database to evaluate MDDs?

**Peer Reviewer Dr. Ingersoll**: I do not agree with the requirement to exceed a MDD in order to identify a sample as toxic, particularly when it is in the context of a SL1/No observed effect level.

**Ecology Response.** One of Ecology's goals in establishing freshwater SQVs was to reflect the same policies and approaches used in establishing chemical and biological criteria for marine sediments in the current SMS rule. This includes setting bioassay endpoints that include an absolute value or response, and statistical difference from control (or reference). Like the marine standards, the SQS level of effects was set as the minimum detectable difference for each of the bioassays. Ecology requested personal experience or knowledge the expert peer reviewers may have, assessing MDD for the proposed bioassays. Ecology recognizes these bioassays are appropriate for protection of the benthic community but do not serve as the sole indicator of toxicity, given that these do not address toxicity from bioaccumulative chemicals, etc. For chemicals that bioaccumulate, the proposed SMS revisions include standards for addressing other ecological receptors and the protection of human health.

**Peer Reviewer Dr. Mount**: For both SL1 and SL2, I think there should be a maximum level of effect that is counted as indicating toxicity regardless of statistical significance. I'm OK with it being both sig diff and a minimum level at the low end, but when effects are large enough, I think sig diff becomes irrelevant. For *Hyalella*, the SL1 criterion for survival is larger (requires more effect) for a 10-d test than for a 28-d test. This makes no sense. I was very surprised that there was no analysis of minimum significant differences within the WDOE database to show what degree of difference was typically needed for those data.

**Ecology Response.** A section was added to the report to discuss the MDD (Appendix C). Regarding the 10 - day versus the 28 - day survival criteria for *Hyalella*, the ASTM round-robin testing indicated slightly higher minimal detectable difference than the long- term test. It may be in part due to moribund organisms in the shorter test that are counted as "live", and which are definitively dead in the longer term test.

**Ecology Question 18.** What are the pros and cons of using two endpoints from the same bioassay (e.g., the Hyalella 28 day test has two endpoints commonly reported from the same test, mortality and growth). Are there ways to maximize the use of the combined results?

**Peer Reviewer Dr. Burton**: Both fine, as one is acute and one is chronic. Hopefully your chronic database will grow.

**Peer Reviewer Dr. Fields**: I recommend using the biomass endpoint to combine mortality and growth, as these endpoints cannot be considered to be independent of each other. Alternatively, using a growth endpoint that represents either mortality or growth effects could be used as a single endpoint.

**Peer Reviewer Dr. Ingersoll**: Add biomass as an endpoint in all Midge and in all Amphipod tests

**Peer Reviewer Dr. Mount**: There is no downside to using both survival and growth data from the same assay, but I would strongly recommend converting to a biomass endpoint which encompasses both effects. This would simultaneously eliminate any awkwardness of having multiple endpoints from a single test.

**Ecology Response:** Data collected from the standard analysis will allow addition of biomass as an endpoint to the bioassays; however, until sufficient data is available, this endpoint can't be incorporated into the biological standards. However, flexibility in the rule language will allow addition of bioassays and endpoints when sufficient data is available to allow development of biological assay standards.

## Additional Comments Made by Peer Reviewers (outside of the above questions Ecology posed)

**Peer Reviewer Dr. Mount**: Remove reference or claims of causality with any particular chemical. Be careful to distinguish "correlation to toxicity" with "cause of toxicity".

**Ecology Response:** The 2011 SQV report has been revised to remove any implied causality.

**Peer Reviewer Dr. Mount**: Low hits to no Hits problematic as false negatives are never very high. Suggests two ways to test if this has adverse impact to the SQVs. Re-run removing the "foundation chemical", to see if other contaminants show up as the primary predictors (if removal impacts other SQVs drastically, then the compound was truly tied to toxicity, if not, it wasn't important and is a "false" driver). Re-run using a 50/50 mix of hit and no-hit values to see if it impacts the final SQVs.

**Ecology Response:** The "foundation chemical" is the issue that caused us to add the "second pass" to the model runs. A description of this has been added to the report, along with an enhanced section on covariance. Additionally, the revised reliability approach in the report addresses the ratio of hits to no-hits.

**Peer Reviewer Dr. Mount**: No accommodation is made for sediment factors that affect availability (AVS, TOC, grain size). Can't ignore the effects of organic carbon on availability but Dave acknowledges there may be a 40-fold effect from OC-normalization vs. a 2500-fold effect from alkylation/bioavailability. Metals also affected by AVS and OC but Dave notes predictive accuracy is unlikely to be improved.

**Ecology Response:** The revised report cites data analyses in previous Ecology evaluations that conclude that TOC normalization does not improve reliability of SQVs. Ecology notes that no other SQV incorporates AVS or TOC. Additionally, if there is a suspicion that availability may be altered due to unusual bioavailability, the SMS rule relies on a numeric chemical standard with a biological override. So the numeric chemical standards can be over-ridden by running bioassays. Anthropogenic TOC sources and unusual TOC (bogs etc) will be specifically mentioned in the rule language as potential reasons why the chemical standards may not be representative.

**Peer Reviewer Dr. Mount**: Are rare toxicants ignored even though they may be major contributors? Does ANOVA screen remove rare toxicants that are major contributors?

**Ecology Response:** The decision was made to develop an SQV for an analyte if the data met a set of minimum requirements (~30 samples). This model has been run on smaller subsets without the ANOVA screen, and the same compounds that are removed by ANOVA tend to have no correlation with toxicity in later stages. Where that toxicity may have been present but not accounted for in developing the SQVs, it will lead to more conservative values for the other analytes but otherwise not affect them. However, where the presence of a chemical without SQVs is known or suspected, project managers should run bioassays (the SMS rule relies on a numeric chemical standard with a biological override).

**Peer Reviewer Dr. Mount**: Acute vs. chronic terminology, Dr. Mount suggests referring to 10 and 28 day, not acute and chronic.

**Ecology Response:** Because definitions of acute and chronic are in sections of the rule that are not being altered in this rule revision, terminology will need to match whatever ends up in the rule language, whether it is acute/chronic, sublethal, etc.

**Peer Reviewer Dr. Mount**: Dr. Mount suggests a wide variety of additional plots, graphs, and maps. Plots and intermediate analyses from FPM to assist in following the method, including derived SQVs vs. overall chemical a distribution plots (interval plots), magnitude of response in contingency tables. Maps with dots diameters indicating the number of samples from the site, and list of primary contaminants at those sites.

**Ecology Response:** Addition of many of these suggestions (interval plots, magnitude of response, etc.) would probably not draw the reader to useful conclusions. There will be chemicals for which the relationship is poorer than others, but it is not possible with this model to just adjust that one without affecting the others. Some new tables and figures have been added, including figure 4-2 (hit/no hit for proposed SQVs) and a map of station locations showing hits and no hits (Figure 2-1). Table 2-4 shows descriptive statistics for chemicals. Attempts to develop maps with all information (number of samples, types of CoCs, magnitude of CoCs, hits, etc.) resulted in either maps that were too difficult to interpret or too many maps.

**Peer Reviewer Dr. Mount**: A number of statements and sentences need more clarity. In particular, Dr. Mount disagreed with the statement that “In general, the freshwater SQVs were developed to protect populations of benthic communities in sediments, rather than individual species, given the wide natural variation in species abundance and richness seasonally and from year to year, especially in freshwater systems”. (Dr. Mounts’ specific comment was: How so? It seems to me that its based entirely on single species.)

**Ecology Response:** This section of the 2011 SQV report was re-worded. The focus was supposed to be that the values are not designed to protect ESA fish or ESA benthic species. Note that the revised report also has better reference to how we dealt with ESA benthic species, which in this region are limited to a group of snails.

**Peer Reviewer Dr. Fields.** This seems like a stretch to say that *Hyalella* and *Chironomus* test endpoints reflect the range of species and life history stages. They could be said to be representative species, but these tests likely were selected because they are standardized, accepted, commonly used tests.

**Ecology Response.** In addition to being standardized tests, they do represent two very different life histories, one being purely aquatic and the other being an aquatic larval form with a terrestrial adult form. Mollusk bioassays are missing, but there are no standardized sediment tests available at this time. The only other really commonly used benthic invertebrates for whole sediment testing are Diporea, another amphipod (pure aquatic life history) and Hexagenia (mayfly, with aquatic larval stage and terrestrial adult phase). Great Lakes also uses a mix of water-phase invertebrates (Daphnids, algae, plants); Washington wanted to stick with organisms with tight links to sediment.

**Peer Reviewer Dr. Fields.** Need to clarify what BPJ calls were based on (selection of SL1 and SL2).

**Ecology Response.** The new report expands on the BPJ calls made for the selection of SL1 and SL2 values.

**Peer Reviewer Dr. Fields.** Need to run validation on the SQGs, not just reliability.

**Ecology Response.** Most SQVs are originally developed without validation, which occurs later once new data are available with which to conduct the evaluation. The reliability evaluation is not intended as a replacement for validation, but rather the best that can be done in the meantime. The reliability evaluation conducted for these SQVs is substantial and rigorous and has resulted in values more reliable than other SQVs. The RSET workgroup discussed whether to set aside a portion of the data set for validation or whether to include it all in the SQV calculations. It was decided to include all the available data in the SQV calculations, and to subsequently validate the SQVs once new regional data became available. Ecology still intends to carry out this process.

**Peer Reviewer Dr. Fields.** Why was Portland 28-day growth results dropped?

**Ecology Response.** It is our understanding that EPA also dropped these values from their final analysis, due to abnormalities in the dataset.

**Peer Reviewer Dr. Fields.** Recommendations for re-runs to address low % hits in dataset- run 50/50 mix and see if it alters SQGs substantially; if it does, then SQGs are too heavily influenced by the large proportion of no hits.

**Ecology Response.** Instead of this extremely work intensive approach, Ecology adopted Burt Shepard's (US EPA) approach of applying statistical evaluations that are independent of the hit/no hit ratio. These new approaches are included in the revised report.

**Peer Reviewer Dr. Fields.** Dr. Fields made recommendations for reliability testing: run on a "per chemical" basis, compare overall performance with respect to the magnitude of the toxicity (are "big hits" being missed, vs. are "minor hits" being missed), use other metrics of reliability.

**Ecology Response.** The SQVs are specifically designed to take into account covariance, additive/synergistic, and bioavailability issues by mimicking how chemicals are actually found in the environment, in mixtures. It will be clearly stated in our regulatory guidance that the SQVs are expected to be applied as a set (and not using quotients or other mathematical manipulations). It should be noted that most SQVs suffer from the opposite problem – they have been developed on an individual basis, and the simple mathematical methods used to combine them are not likely to accurately reflect the complex interactions between chemicals in the environment.

Ecology adopted Burt Shepard's approach of applying statistical evaluations that are independent of the hit/no hit ratio. These new approaches are included in the revised report.

**Peer Reviewer Dr. Fields.** Is difference between SL1 and SL2 meaningful?

**Ecology Response.** This is a two part question- is the difference meaningful in a biological sense, and from an analytical chemistry perspective. From a biological perspective, it represents the next most sensitive biological response from all the assessed endpoints, and as such, should be considered meaningful. From an analytical chemistry perspective, we re-analyzed the proposed SLs to ensure that there was at least a 20% relative percent difference between the SQS (SL1) and CSL (SL2) values; this resulted in changing the CSL (SL2) for chromium and di-n-butyl phthalate. A new section was added to the revised report that addresses this question.

**Peer Reviewer Dr. Fields.** Suggests more analyses, tables and maps, including:

- Distributions by endpoints (concentration ranges of detected data, max non-toxic concentration, number of toxic samples).
- Something to indicate the dominance of certain projects (e.g. Portland) for certain endpoints/chemicals. See App A. table
- Plot magnitude of toxicity vs. concentration of chemical/chemical's SQV

**Ecology Response.** Addition of many of these suggestions (distribution by endpoints) would probably not draw the reader to useful conclusions. There will be chemicals for which the relationship is poorer than others, but it is not possible with this model to just adjust that one without affecting the others. Some new tables and figures have been added, including figure 4-2 (hit/no hit for proposed SQVs) and a map of station locations showing hits and no hits (Figure 2-1). Table 2-4 shows descriptive statistics for chemicals. Attempts to develop maps with all information (number of samples, types of CoCs, magnitude of CoCs, hits, etc.) resulted in either maps that were too difficult to interpret or too many maps.

**Peer Reviewer Dr. Ingersoll.** Insufficient information has been provided in the report describing the process used to develop the draft WA-OR-ID sediment quality values (SQVs). Specifically, the report needs to be expanded to summarize: (1) primary toxicity or chemistry data used to develop the database and to derive the SQVs; and, (2) results of various initial analyses of the database. Without this summary information, I cannot provide a detailed review of the proposed procedure for developing regional SQVs for WA-OR-ID. At a minimum, the report needs to provide tables summarizing:

a. **Number of samples included in the database by region within each of the three states.** Included in this table, there needs to be a summary of the percent of the samples that are toxic by endpoint that are: (1) significantly different from control and (2) are significantly different from control and by a reduced percentage from control (e.g., <10%, <20%, <30%).

**Ecology Response.** Data on region and chemical families have been added to the revised report in Appendix A. Only samples determined to be a "hit" are statistically different (see added section on minimum detectable differences in Appendix C1). Because of the way the model works, and framework that Ecology uses, Ecology does not believe that the magnitude of response is useful in the context of the report and the developed SQVs.

b. **Number of samples excluded from the database by region within each of the three states.** Included in this table, there needs to be a summary of the number of stations that were screened out of the database across all of the regions and the associated rationale for data exclusion (see Section 2.1.2).

**Ecology Response.** See Appendix B of the 2011 SQV report for additional information.

c. **Regional reliability of SQVs.** Section 3 of provides summaries of reliability estimates across the database by toxicity endpoint. There is a need to determine how well the SQVs estimate toxicity on a regional basis within each of the three states. See for example the summaries provided in Tables 3, 4, and 5 of Ingersoll et al. (2001; Ingersoll et al. 2001. Predictions of sediment toxicity using consensus-based freshwater sediment quality guidelines. Arch Environ Contam Toxicol 41:8-21).

**Ecology Response.** See Appendix D2 for additional information; also note that no data is available for Idaho.

d. **Existing SQGs included in Table 3-3 and 3-4.** A table summarizing these existing SQGs referenced in Table 3-3 and in Table 3-4 needs to be included in the report.

**Ecology Response.** The 2011 report builds on the work initiated in Phases 1 and 2, and the other SQVs are discussed in the documents from 2002 and 2003 (2002 and 2003, SAIC and Avocet) .

e. **Comparisons of existing SQGs to the proposed SQVs**. A table is needed comparing the relative differences in existing SQGs compared to the proposed SQVs. While some similarities are evident between some of the existing SQGs and proposed SQVs (e.g. Hg PEC of 1.06 µg/g vs. Hg SL1 of 0.66 µg/g), there are also marked differences (e.g., zinc PEC of 459 µg/g vs. Zn SL1 of 3200 µg/g and Zn SL1 of >4200 µg/g). These comparisons should be presented discussed in the report. Moreover, this table would identify chemicals of concern for which there are proposed SQVs with no existing SQGs (e.g., Tributyltin).

**Ecology Response.** While we agree that a table comparing the SQV sets would highlight where the lists of chemicals with SQVs differ, we disagree that a comparison of number to number values is useful. Rather, Ecology prefers to evaluate the performance of the SQV set on predicting toxicity, not differences between the individual values.

The FPM SQVs are specifically designed to take into account covariance, additive/synergistic, and bioavailability issues by mimicking how chemicals are actually found in the environment, in mixtures. It will be clearly stated in our regulatory guidance that the SQVs are expected to be applied as a set (and not using quotients or other mathematical manipulations). It should be noted that most SQVs suffer from the opposite problem – they have been developed on an individual basis, and the simple mathematical methods used to combine them are not likely to accurately reflect the complex interactions between chemicals in the environment.

**Peer Reviewer Dr. Ingersoll.** The SQVs approach has been developed using a floating percentile method (FPM). The FPM is based on the evaluating the frequency of SQV exceedances or the frequency of toxicity of samples within the database. However, the FPM as described in the report does not address the magnitude of contamination or the magnitude of toxicity. Samples are placed in one of two toxicity categories: toxic or not toxic and samples are placed in one of two chemistry categories: below an SQV or above an SQV. Over the past decade, there have been substantial advances in development and application of sediment quality guidelines (SQGs) that incorporate the magnitude of toxicity (e.g., percent survival) and the magnitude of chemistry (e.g., mean quotients) in the context assessing sediment contamination. Wenning et al. (2005) provides examples of empirical and mechanistic SQG approaches for evaluating magnitude of toxicity or magnitude of chemical concentration. Importantly, the proposed SQV approach needs to be revised to provide mechanisms for addressing: (1) the magnitude of chemical exceedances of SQGs and (2) the magnitude of toxicity. Wenning RJ, Batley G, Ingersoll CG, Moore DW, editors. 2005. Use of sediment quality guidelines and related tools for the assessment of contaminated sediments. Pensacola FL: SETAC Press, 783 p.

**Ecology Response.** Magnitude of contamination and toxicity are part of risk assessment evaluations; these values were designed to function as risk management screening values. The SQS and CSL levels of effects do identify the lower range of effects that serve as triggers for cleanup but additional considerations can be used for determining the scope of cleanup actions. Once a site has been listed as an area of concern, then Ecology site managers can address magnitude of chemical and toxicity exceedances using appropriate methodologies including those mentioned in this comment.

**Peer Reviewer Dr. Ingersoll.** The SQV approach based on the FPM does not adequately address the influence of mixtures or the contribution of individual chemicals of concern on observed sediment toxicity. Equal weight is given to an exceedance of a major class of compounds (e.g., total PAHs) relative to a single exceedance of a chemical (e.g., carbazole). Wenning et al. (2005) provides examples of approaches for evaluating the influence of mixtures and the influence of individual chemicals of concern associated with application of empirical or mechanistic SQGs. Importantly, the proposed SQV approach needs to be revised to provide mechanisms for more directly addressing the influence of mixtures on the predicted toxicity. The proposed SQV approach described in the report also needs to be revised to provide mechanisms for identifying the influence of individual chemicals of concern on the observed toxicity.

**Ecology Response.** The FPM SQVs are specifically designed to take into account covariance, additive/synergistic, and bioavailability issues by mimicking how chemicals are actually found in the environment, in mixtures. It will be clearly stated in our regulatory guidance that the SQVs are expected to be applied as a set (and not using quotients or other mathematical manipulations). It should be noted that most SQVs suffer from the opposite problem – they have been developed on an individual basis, and the simple mathematical methods used to combine them are not likely to accurately reflect the complex interactions between chemicals in the environment.

Mechanisms are important when summing chemical classes, and we did take that into account when developing summed CoCs. However, because the FPM approach actually does take co-variance and mixtures into account in an empirical manner, it really doesn't matter what the mechanism is once chemicals have been appropriately summed.

**Peer Reviewer Dr. Ingersoll.** A primary objective of the report was to improve on limitations of existing SQGs (e.g., page ES1, 1st paragraph). While an evaluation of the reliability of existing SQGs was included in Tables 3-3 and 3-4, no general summary of these existing SQGs has been provided in the report, and the primary literature describing these existing SQGs has not even been cited in the report. Moreover, it is not clear as to the source of the ERLs and ERMs (marine or freshwater?). The report needs to be expanded to provide an overview of the narrative intent of existing SQGs referenced in Section 3. The report also needs to summarize these actual SQG values in an appendix to the report. Importantly, the reader needs to understand the source and narrative intent of the existing SQGs and reader needs to know how different the existing SQGs are compared to the proposed SQVs.

**Ecology Response.** When Ecology first embarked on development of freshwater SQVs, each of these existing SQV sets was evaluated to determine how well they predicted toxicity in Washington and Oregon sediments. The results showed very poor reliability in predicting the toxicity of freshwater sediments in the Pacific Northwest region (SAIC and Avocet 2002). This evaluation was repeated in 2008, with the same results (Avocet 2010). This evaluation is what prompted the agencies to consider development of an alternative approach to setting SQVs that would be more predictive. As was done for the SMS marine standards, Ecology ensured regional sediment samples were collected and analyzed using both Ecology and ASTM approved bioassays. Our goal all along has been to improve on the reliability of existing SQV sets. If at any time we had found that this was not the case, the development effort would have been discontinued in favor of the existing SQVs.

Additional language has been added to the report regarding an overview of the intent of these SQVs; note that the actual SQVs are summarized in the Executive Summary, not in an appendix. The report also was modified to clarify that all ERLs and ERMs were freshwater values.

**Peer Reviewer Dr. Ingersoll.** Tables 3-3 and 3-4. These two tables summarize the frequency of reliability (percentages) across different metrics. A critical piece of information missing from each of these reliability estimates is the number of samples that are used to estimate each of these frequencies. It appears from the data summarized in Table 3-2, there are a limited number of toxic samples for many of the endpoints (less than 15% across most of the endpoints?). If there are a limited number of toxic samples available for estimating false positive or false negative rates of SQVs, this limited number of toxic samples limits the utility of the reliability estimates (e.g., dramatic changes in reliability estimates can occur when there are a limited number of toxic samples within a category of interest). That said, only with sample number identified in Tables 3-3 and 3-4, can one then understand how robust particular measures of reliability might be across toxicity endpoints.

**Ecology Response.** New figures have been added to provide this information (See figure 4-2 series). New reliability analyses have also been conducted using metrics that are independent of the hit/no-hit ratios (Section 4.3).

**Peer Reviewer Dr. Ingersoll.** Tables 3-3 and 3-4. Comparisons are made between reliability of the proposed SQVs and existing SQGs. It is not clear as to how these estimates of reliability were made to the existing SQGs. Moreover, the number of toxic samples used to make each of these comparisons for each endpoint is not included in these tables (see comment above). That said, for the existing threshold-type SQGs (e.g., ERLs, TECs in Table 3-3), the frequency of false positives, are expected to be high (not intended to be predictive of toxicity). These threshold SQGs are to have a low false negative rate which is typically the case in Table 3-3. Similarly, for the probable-type thresholds (e.g., ERMs, PECs in Table 3-4), the frequency of false negatives is expected to be high (not intended to be predictive of the lack of toxicity). What is surprising is the frequency of false positives is also relatively high for the false positives using the existing probable threshold. We have found the false positive rate to be relatively low when using probably-type SQGs to predict toxicity. Importantly, a description is needed in the report as to how the false positive and false negative rates were estimated for the existing threshold SQGs in Table 3-3 and for the exiting probable thresholds in Table 3-4 before any definitive conclusions can be made on the reported reliability of these existing SQGs in Section 3 of the report. Moreover, the report needs to cite and discuss the primary literature regarding the reliability estimates of existing SQGs compared to the reliability estimates of existing SQGs reported in Tables 3-3 and 3-4.

**Ecology Response.** Language has been added to clarify comparisons and the evaluation of false negatives and positives, and new reliability evaluations have been added to the report based on comments from Burt Shepard, EPA.

**Peer Reviewer Dr. Ingersoll.** Tables 3-3 and 3-4. Overall reliability estimates the FPM values are provided in Tables 3-3 and 3-4. What is missing from the report is an evaluation of the reliability of individual SQV. It is likely that many of the individual SQVs listed in Table ES-1 would exhibit a wide range in reliability. This is a critical analysis that needs to be included in the report, particularly when individual SQVs will be used to evaluate chemicals of interest.

**Ecology Response.** The FPM SQVs are specifically designed to take into account covariance, additive/synergistic, and bioavailability issues by mimicking how chemicals are actually found in the environment, in mixtures. These empirically derived SQGs are not designed to be used on individual chemicals. It will be clearly stated in our regulatory guidance that the SQVs are expected to be applied as a set (and not using quotients or other mathematical manipulations).

**Peer Reviewer Dr. Ingersoll.** Narrative intent of SQS and SL1 values. The SL1value (and SQS) is intended to be a no observable effect level (e.g., Section 2.6.1). Given this narrative intend of an SL1 value, it is surprising that a false negative rate of 19 to 20% (Table 3-3) would meet the reliability goal of a no observable effect level. From my perspective, misclassifying 1 of every 5 samples, is not an acceptable definition of a no observable effect concentration, particularly in a database with a <20% incidence in toxicity to begin with.

**Ecology Response.** The SMS rule was not designed to require proof of absence of toxicity. Rather the SMS rule was developed to manage environmental risk by minimizing adverse effects to a population or community of benthic organisms. WAC 173-204-315, “no adverse effects” for marine bioassay tests are defined using effects thresholds of 15-30% compared to reference, depending on the test endpoint. These thresholds are based on the minimum detectable difference in these bioassay endpoints.

The freshwater SQVs were designed with the same principles in mind, except they are compared to the control, a more stringent standard. Therefore, 15-20% difference from control is well within the regulatory definition of “no adverse effects” for bioassay tests, and is based on minimum detectable difference, or the point at which an adverse effect can first be observed. Effects below this level are considered “no adverse effects,” while effects exceeding this level are considered “minor adverse effects” up to the higher CLS/SL2. As a result, the primary goal stated above is met, as there are not expected to be biologically observable or statistically meaningful adverse effects below this level.

The regulatory framework that the current SQVs would be applied within differs somewhat in that it requires strong defensibility of the SQVs at the SL1 level, not only in predicting areas that are non-toxic, but in accurately predicting toxicity, so that both agency and regulated community resources are not wasted on unnecessary characterization efforts. The SL1 levels are used not only to screen sediments out, but to determine which sediments may reasonably be expected to be toxic and thus warrant further attention.

Ecology has emphasized the importance of accurately predicting both the presence and absence of toxicity (reliability) in developing the proposed SQVs and, in the SQV report, compared reliability to other SQVs. The application of the other SQVs in the regulatory context was also examined in the draft EIS. These efforts determined the current SQVs are more defensible with respect to ability to accurately predict presence or absence of toxicity. In the regulatory arena they result in a more efficient focus of effort and expenditures to identify and characterize sites.

The TEL/TEC values have an 85-95% false positive rate, which indicates that almost no sediments would be screened out, eliminating the value of having any screening levels. While TEL/TEC values represent a conservative approach, Ecology must make risk management decisions based on accurate predictions that sediment is contaminated in order to require further costly analysis and cleanup. Using TEL/TEC type values would be no more effective than Ecology’s current practice, requiring bioassays at all stations for all projects. However, this approach is costly and does not provide predictability and consistency to make cleanup decisions. Alternatively, more accurate SQVs can and have been developed, and are proposed here. This was further discussed in the draft EIS which compares the pro and cons of adopting different SQVs.

A distinction needs to be made between the SL1 interpretive guideline and the acceptable error rate – which are two entirely different things. The SQS or SL1 guideline (no observable effects level) is based on the protection of the overall diversity and functions of the benthic community. This means protection of benthic organisms at the population level. This is applied in the regulatory arena and is not intended to represent the most sensitive indicators of toxicity which might be more useful in risk assessment rather than risk management.

**Peer Reviewer Dr. Ingersoll.** False negative error rate. It is my understanding that a decision needs to be made before the FPM can even be run is to pre-determine an acceptable false negative rate. This decision pre-defines other measures of reliability in addition to the false negative error rate, and thus makes FPM approach more of a risk management tool rather than a risk assessment tool. Setting an allowable false negative error rate of 20%, and then allowing the FPM to increase false positive rates is not a conservative approach for protecting the environment. A maximum false negative rate of 5% or 10% would be more in line with the designed level of protectiveness of other ecological benchmarks, such as the USEPA aquatic life criteria or other types of existing SQGs.

**Ecology Response.** These values are meant to be used as a risk management tool, not a risk assessment tool. Additionally, there seems to be some confusion regarding the SL1 interpretive guideline with the acceptable error rate – which are two entirely different things. The SQS or SL1 guideline (no observable effects level) is based on the protection of the overall diversity and functions of the benthic community. This means protection of benthic organisms at the population level. This is applied in the regulatory arena and is not intended to represent the most sensitive indicators of toxicity which might be more useful in risk assessment rather than risk management. Also, these are only endpoint-specific error rates for individual stations, which will be much lower once combined and with multiple stations of data on which to base decisions.

**Peer Reviewer Dr. Ingersoll.** Independent evaluation of predictive ability of the proposed SQVs. Comparisons between the FPM and other SQGs need to be conducted using an independent dataset. The evaluations of reliability in the report focused on evaluation using the database that was used to derive the SQVs, which may result in higher reliability of the proposed SQVs compared to using the existing SQGs to predict toxicity in the database. Since this independent predictive validation of the SQVs has not been performed in the report, the utility of the proposed SQVs compared to existing SQGs cannot be determined.

**Ecology Response.** Most SQVs are originally developed without validation, which occurs later once new data are available with which to conduct the evaluation. The reliability evaluation is not intended as a replacement for validation, but rather the best that can be done in the meantime. The reliability evaluation conducted for these SQVs is substantial and rigorous and has resulted in values more reliable than other SQVs. The RSET workgroup discussed whether to set aside a portion of the data set for validation or whether to include it all in the SQV calculations. It was decided to include all the available data in the SQV calculations, and to subsequently validate the SQVs once new regional data became available. Ecology still intends to carry out this process.

In the freshwater environments that exist in the Pacific Northwest, ranging from small mountain streams to large regional rivers, from dammed areas to channeled agricultural ditches, and from acidic lakes and wetlands to alkaline lakes and mining-impacted watersheds, it would be difficult to conduct validation against benthic community data. Most of these surveys do not have acceptable reference areas, and a 4-year RSET process to attempt to identify appropriate reference areas in the region largely failed. This type of evaluation (e.g., Ingersoll 2001) is really only possible in freshwater areas such as the Great Lakes, which are large and relatively uniform in their benthic environments and communities, more like Puget Sound.

**Peer Reviewer Dr. Ingersoll.** Page ES-1, 2nd paragraph from the bottom. The statement is made that all data was collected using ASTM-approved or Ecology-approved bioassay methods and chemistry method. A table needs to be included in the report that summarizes the key test acceptability requirements from ASTM or from Ecology that were used to judge acceptability of the regional sediment toxicity data (see page 5, 2nd paragraph). The only acceptability requirement mentioned in the report was in reference to the number of replicates tested (5 replicates was established as the acceptability criterion in Section 2.1.2 [even though ASTM E1706 states 4 replicates is acceptable; see Section 15.2.2.4 in ASTM E1706]).

**Ecology Response.** QA requirements are exhaustive and too long to be listed in the report. However, the methods and protocols used for QA2 screening of chemistry and bioassay data have been more thoroughly described and referenced in the revised report.

**Peer Reviewer Dr. Ingersoll.** Page 8, 2nd paragraph. Reference is made to a “biological over-ride” of SQVs by individual toxicity tests. I do not think results of individual toxicity tests should be used in such a powerful way. Importantly, there is a need to assess sediment contamination using a “weight of evidence” approach (not trumping one robust endpoint with limited data generated from a potentially confounded single toxicity endpoint). Specifically, there should be a mechanism in the WA-OR-ID sediment assessment framework to go beyond initial screening of sediment toxicity using: (1) empirical SQVs and (2) sediment toxicity tests. For example, the use of: (1) mechanistically based SQGs (equilibrium partitioning) and (2) toxicity identification evaluations should be included in the WA-OR-ID sediment assessment program. Use of tools beyond toxicity tests and SQVs are needed when costly decisions are to be made regarding actions such as cleanup of sediment or source control. See Wenning et al. (2005) for detailed discussions of the use of weight of evidence approaches for assessing potential risks of contaminants in sediment.

**Ecology Response.** The proposed bioassay over ride is consistent with the existing SMS rule. Note that bioassay over ride requires use of multiple bioassays, not a single endpoint.

**Peer Reviewer Dr. Ingersoll.** Page 8, 3rd paragraph. The statement is made that “identification of adverse biological effects generally involves a statistical difference from control or reference plus some threshold of effects.” This statement is not supported by either guidance or by recommendations in either ASTM or in USEPA methods for establishing toxicity of sediments. Requiring a percentage difference from control or reference is a policy judgment that some regional groups have used to identify toxicity beyond comparisons to control sediments or reference sediments.

**Ecology Response.** These references provide only protocols for running bioassays, not guidance on how to interpret the data. Text has been amended by adding "In Washington State sediment programs," to the beginning of the sentence for clarification.

**Peer Reviewer Dr. Ingersoll.** Table 2-2. Analyses are needed in the report that evaluating reliability of the SQVs relative to the various procedures used to establish toxicity relative to control conditions (e.g., significance from control alone versus significance from control and a percent difference from control).

**Ecology Response.** New Appendix C added to address this comment.

**Peer Reviewer Dr. Ingersoll.** Table 2-2. I do not understand the rationale for establishing separate quality assurance for response of organisms in control sediment versus reference sediment (e.g., Hyalella azteca 10-d testing <20% mortality required for control sediment and <25% mortality required for reference sediment). Why are these requirements not the same? What makes 25% mortality acceptable in a reference sediment?

**Ecology Response.** Control represents the best case scenario most closely imitating an organisms normal habitat conditions and is expected to have a low level of effects on survival. Reference is intended to reflect physiochemical conditions similar to test sediments but without effects of toxic substances. Extensive evaluations of program data (both marine and freshwater) have shown that reference samples generally have equal or higher toxicity than control samples.

On a programmatic basis, the average of reference samples is higher than control samples. In addition, reference samples (field samples) tend to have higher variability among replicates than control samples. Therefore, decreasing the statistical power of the comparison and increasing the minimum detectable difference. For these reasons, comparison to control is more conservative than comparison to reference, in general (though individual sites may vary). Because we are attempting to maintain program consistency between the SMS marine and freshwater standards, the difference is worth noting. It would be inappropriate to have both a more conservative comparison framework and more conservative thresholds for freshwater versus marine. This change makes them more consistent, though that was not its purpose.

**Peer Reviewer Dr. Ingersoll.** Table 2-2. I do not understand why the required differences from control sediment are different from the required differences from reference sediment (e.g. Hyalella azteca 10-d testing 15% for control and 25% for reference). Why are these requirement not the same? What makes at a 15% difference from a control a problem compared to a 25% difference from reference a problem?

**Ecology Response.** These are not interpretive criteria, they are a QA performance limit required for determining if a test is valid.

**Peer Reviewer Dr. Ingersoll.** Table 2-2. The rationale for requiring a 20 to 40% difference in growth from control to establish toxicity in amphipods or in midge toxicity tests is not justified in the report. These are a very large difference from control (e.g., toxic samples may be identified as non-toxic using these broad growth ranges).

**Ecology Response.** The values mentioned in this comment are for the CSL value (upper screening value), not SQS value (lower screening value). New Appendix C discussion on minimal detectable difference was added to address this comment.

**Peer Reviewer Dr. Ingersoll.** Page 8, last paragraph to page 11, 2nd paragraph. I am completely lost by the rationale that has been used come to the following type statements for each species and each endpoint under each of the subsections on pages 8 to 11: "Given this, the maximum mortality that would observed at the SQS/SL1 level would be 30-35%...". The report needs to describe in much more detail regarding the rationale for each of these "given statements" on pages 8 to 11 for all species and endpoints.

**Ecology Response.** New Appendix C added to address this comment.

**Peer Reviewer Dr. Ingersoll.** Page 10, 1st paragraph. "Policy" objectives should not be used to establish recommended toxicity endpoints. See also Page 16, 2nd paragraph (error levels were set at comfort levels?).

**Ecology Response.** New Appendix C added to address this comment. Language inferring to "comfort levels" when discussing the goals set by the agencies has been revised.

**Peer Reviewer Dr. Ingersoll.** Section 2.4 Exploratory Analyses and Section 2.5 Final Model Runs (page 16 to 18). The report needs to summarize the results of these exploratory and final model runs. Without these summaries, the acceptability of the decisions in Section 2.4 and in 2.5 regarding the final modeling effort cannot be reviewed.

**Ecology Response.** New Appendix D added to address this comment.

**Peer Reviewer Dr. Ingersoll.** Use of ANOVA to eliminate chemicals. The FPM uses an ANOVA test between hit and no hit distributions to justify the chemical lists included in and excluded from the model. The ANOVA test has an assumption of normality and may be invalid for most of the chemical distributions sediments. Most of the distributions of chemicals in sediment are highly skewed. Re-analyses are needed to determine if the chemicals eliminated exhibit statistical differences between hit and no hit concentrations using non-parametric tests.

**Ecology Response.** This statistical argument is not without merit from a theoretical standpoint. However, it is just a screening approach that has been used for 12 years with this model and shown to work. Even if the distributions aren't strictly as they should be, it is the relative magnitude of the association that is important and provides useful information (e.g., no association for any endpoint used to screen out chemicals).

It is not possible to use nonparametric statistical tools with the spreadsheet method we have. Therefore to do something different we'd have to go to R or SPSS or something equivalent, somewhat obviates the point of a simple-to-use method.

**Peer Reviewer Dr. Ingersoll.** Page ES-2, 2nd to last paragraph. The conclusion that the SQVs are expected to be protective of endangered species in Washington, Oregon, or Idaho has not been adequately addressed in the report. Many species of concern (e.g., lamprey in Portland Harbor, white sturgeon and mussels in the Upper Columbia River, snails in Idaho, to name a few) may be highly sensitive to contaminants of concern in sediment across the three state region. Even if there are no listed species in WA, OR, or ID that are present in areas where dredging or clean up is likely (page 18, last paragraph), the SQVs will likely be applied on a regional bases to assess sediment contamination where there may be listed species (e.g. White sturgeon inhabiting metal-contaminated sites in the Upper Columbia River).

**Ecology Response.** These values are meant to be protective of benthic invertebrates, not fish. Fish must be dealt with on a site specific manner following the narrative ecological standard in the proposed SMS rule. Note that no SLs are available for the protection of fish for most compounds. Regarding the ESA species, NOAA and USF&W were directly involved in the RSET workgroup and were tasked with determining if individual ESA-listed species were protected. They confirmed there were no endangered benthic species in these regions.

**Peer Reviewer Dr. Ingersoll.** Page i. A regional sediment evaluation team (RSET) provided guidance on the initial development of the SQVs. Has the draft report should be reviewed by each member of the RSET team? It is surprising that USFWS members of the RSET would agree with the conclusion that the SQVs are protective of endangered species.

**Ecology Response.** The RSET agencies did review the draft report. See response above regarding USFWS review.

**Peer Reviewer Dr. Ingersoll.** The development of regional SQVs for WA, OR, and ID would benefit from organizing a workshop including a national sediment evaluation team to further discuss methods for (1) developing and applying both empirical and mechanistic SQGs and (2) conducting sediment toxicity assessments.

**Ecology Response.** Local and national experts were involved throughout the development and review of the regional database and the FPM approach for developing SQVs. The development process included experts from Washington State Department of Ecology, Department of Natural Resources, EPA, Corps of Engineers, USF&WS, NMFS, Oregon DEQ, Idaho DEQ and nationally recognized experts from the private sector. The extensive review by the Science Panel, the Sediment Work Group, and this external Scientific Peer Review group included national and regional experts representing academia, USGS, EPA, NMFS, ports, private sector, local and municipal governments and tribes.

**Peer Reviewer Dr. Ingersoll.** Title: It would be more appropriate to re-title the report "**Options for establishing** benthic SQVs for freshwater sediments in WA, OR, and ID", based on the statement made on page 28, last paragraph, that this report is intending to provide agencies with options for selecting SQVs, not with absolute SQVs.

**Ecology Response.** Comment noted. The stated objective for this project is captured in the title, that is to develop benthic SQVs using data from the region. The reported SQVs do represent an option available to the individual states to use as each chooses.

**Peer Reviewer Dr. Ingersoll.** Page ES-1, 2nd paragraph. "Sub chronic" is not a routinely used term (e.g., ASTM or in USEPA methods focus more on duration and endpoints). Moreover, calling a 28-d lethal endpoint an "acute" endpoint is not a routinely used term (e.g., Page 7, last sentence).

**Ecology Response.** Ecology is working on terminology that is consistent with the definitions in the existing rule, as well as consistent with the intent of the new rule language being proposed. The intent is to differentiate between mortality and sublethal effects, as well as between short (10 day) and longer exposure time (20, 28 day) bioassays.

**Peer Reviewer Dr. Ingersoll.** Page ES-2. The conclusions on accuracy of the proposed SQVs and the comparisons of the proposed SQVs to existing SQG has not be adequately supported by the analyses presented in Table 3-3 and in Table 3-4

**Ecology Response.** New sections added on both comparison and reliability.

**Peer Reviewer Dr. Ingersoll.** Table ES1-1. Why is chlordane not listed?

**Ecology Response.** Chlordanes (total) were eliminated based on SQVs that would be higher than the greatest reported concentrations of these compounds. See Appendix B for this and other CoCs that were screened out. Note that if a screened out compound is a known CoC's at a site, Table 2-4 has a list of the maximum concentration reported in data; if a site exceeds these concentrations, then bioassays must be conducted.

**Peer Reviewer Dr. Ingersoll.** Section 1.3. What national sediment experts were asked to review the SQVs developed using the FPM?

**Ecology Response.** Local and nationally recognized sediment experts who reviewed the SQVs included Alan Burton, Jay Fields, Chris Ingersoll, Dave Mount, Jack Word, Clay Patmont, Joanne Snarski, Burt Shepard, Pete Rude, Glen St. Amant, Paul Fuglevand, Teri Floyd, Bruce Duncan, Mike Riley.

**Peer Reviewer Dr. Ingersoll.** Section 2.1.1. Cite Appendix B in this first paragraph.

**Ecology Response.** Done.

**Peer Reviewer Dr. Ingersoll.** Section 2.1.2. Summarize the number of stations that were screened out of the database across all of the regions.

**Ecology Response.** Information was included in Appendix B.

**Peer Reviewer Dr. Ingersoll.** Page 5, 2nd paragraph. That data are "too old" is not a test acceptability criterion described in ASTM standards. Is such a criterion described in the Washington Department of Ecology standards?

**Ecology Response.** The RSET and the DMMP groups developed the "recency of data" evaluation based on changes in analytical chemical and bioassay methodology. Additionally, older data QA documentation is difficult/expensive to obtain, and QA2 level is required for data supporting the rule.

**Peer Reviewer Dr. Ingersoll.** Page 5, 3rd paragraph. Explain the rationale for requiring a minimum of 30 detected values was chosen as a requirement to include a chemical in the derivation of an SQV.

**Ecology Response.** This was a minimum requirement established by consensus among sediments experts working in the northwest to develop both marine and freshwater SQVs. It provides a reasonable expectation that an adequate range in both chemical concentration and bioassay effects will be represented with enough samples to adequately address variability.

**Peer Reviewer Dr. Ingersoll.** Section 2.1.3, 1st paragraph. The statement is made that organic carbon normalization was not done because it is difficult to understand by the regulated community. This is not a sufficient rationale for not including an evaluation of the influence of organic carbon normalization on the reliability of the proposed SQVs. If this rationale is used to exclude organic carbon normalization procedures, then one would be hard pressed to justify the use of the FPM, given how difficult the procedure is to understand.

**Ecology Response.** Bioavailability of some compounds is affected by OC but the amount and nature of OC is so highly variable to include OC normalization would not be defensible. OC-normalizing would be expected to reduce apparent toxicity where OC is higher. Worst case non-OC normalized values would be (more sensitive) more conservative. Additionally, no other SQV takes OC into account. And in cases where OC is very low, OC normalization causes exceedances based on non-detects on a regular basis- it is agreed upon at Ecology that OC normalization is not prudent when OC is low (0.5 to 0.8%TOC). Thus, it would require the use of TWO sets of standards, one where OC is high, one where OC is low, with little to no benefit to predictiveness.

TOC-normalization was tested side-by-side with dry weight values both in 2003 and in 2008, as well as for a number of other projects, listed in Avocet (2010). These evaluations did not improve the reliability of the results. The marine AETs are the only state or provincial values we are aware of that are TOC-normalized, and it was recognized at the time they were developed that the dry weight AETs were equally predictive. TOC-normalization of the marine AETs was largely done in deference to theory, but it has not been demonstrated to improve predictiveness. In addition, there have been significant implementation difficulties associated with the OC-normalized values, including sediments that had TOC outside the equilibrium partitioning range (too low or too high), sediments with excessive TOC of anthropogenic origin, and general difficulties explaining the approach to the regulated public.

A number of other summing and normalization approaches were also tested, with the result that some chemical classes were summed. Other approaches did not improve reliability.

Bioavailability will always differ between sediment samples due to the origin and nature of the chemicals themselves and the many different physical and chemical aspects of the sediment and porewater. A considerable effort has been focused on better understanding the contribution of some of these different factors, such as OC. Ultimately, if there were clear evidence that these other ways of looking at the chemistry served to consistently improve the predictive accuracy of the SQVs, then we would adopt a different approach. The FPM is the only method for developing SQVs that explicitly deals with the bioavailability of a chemical in a synoptic data set by managing its effect on false negatives and false positives. Also, our regulatory framework with a range between the lower and upper regulatory levels and a biological override allows appropriate consideration of site-specific factors where there is reason to believe the SQVs may be less effective predictors.

**Peer Reviewer Dr. Ingersoll.** Section 2.1.3, 1st paragraph. The statement is made that organic carbon analyses of the limited 2002 data did not improve reliability. These types of analyses need to be done on the expanded 2010 database, given the limitations that were identified in the 2002 database.

**Ecology Response.** See response above. Additionally, the RSET workgroup decided not to do this, given that there were many other examples of this analysis besides the 2002 report and that no other agency programs are using it. It was agreed that we want to move to dry weight overall. Therefore, it was not considered a priority to redo this analysis, which is very labor-intensive.

**Peer Reviewer Dr. Ingersoll.** Section 2.1.3, 3rd paragraph. Summarize the analyses comparing comparisons of THP and PAHs relative to explaining petroleum toxicity, particularly for those stations where TPHs were not measured.

**Ecology Response.** See new Appendix D1.

**Peer Reviewer Dr. Ingersoll.** Section 2.1.4, 1st paragraph. The report has not adequately summarized the analyses leading to the conclusion that there “appears” to be no reliability advantage to estimating in the 2002 database when making toxicity designations relative to control versus relative to reference conditions. Moreover, these analyses need to be performed using the expanded 2010 database, given the limitations identified in the report regarding the 2002 database.

**Ecology Response.** See new Appendix D3.

**Peer Reviewer Dr. Ingersoll.** Section 2.1.4, 1st paragraph. How would one “standardize” reference areas?

**Ecology Response.** Revised version now cites the RSET and DMMP paper on how to develop “reference areas”, based on the Portland effort. A set of minimum requirements must be satisfied. Universally this has been agreed to as limiting.

**Peer Reviewer Dr. Ingersoll.** Table 2-2, footnote a. This footnote is applicable to all of the QA control conditions listed in Table 2-2.

**Ecology Response.** Change made.

**Peer Reviewer Dr. Ingersoll.** Page 10, 4th paragraph (and throughout the report). Ingersoll et al. (2008) did not make these “recommendations’, these were “suggestions”.

**Ecology Response.** Change made.

**Peer Reviewer Dr. Ingersoll.** Section 2.1.6. Screening of data using ANOVA. Insufficient information has been provided in the report to justify the results of these screening analyses (and Appendix B does not provide enough “detail”).

**Ecology Response.** Appendix B was revised and expanded. However, in order to really obtain the details, one would need to go through the actual model spreadsheets. Complete tables have been provided showing the levels of significance that result, not sure what more could be provided.

**Peer Reviewer Dr. Ingersoll.** Page 13, last paragraph. Insufficient information has been provided in the report describing how “optimized’ site-specific SQVs can be developed from the SQVs summarized in Table E1-1.

**Ecology Response.** It is the FPM model that can be used to develop “optimized” site-specific SQVs but this requires a synoptic data set of adequate size (comprised of the spectrum of CoCs present and a suite of bioassays).

**Peer Reviewer Dr. Ingersoll.** Table 3-7. The floating percentiles in each column need to be identified relative to the specific toxicity endpoint.

**Ecology Response.** Change made.

**Peer Reviewer Dr. Ingersoll.** Reference. Most all of the references are not complete (volume numbers missing, presentations but not publications cited). ASTM 2005 should be cited as ASTM 2010. USEPA (2009) is not a correct citation (it is USEPA 2003).

**Ecology Response.** This didn’t exist when this work was conducted.

**Peer Reviewer Dr. Ingersoll.** Appendix A. This table needs to be expanded (see Comment 1)

**Ecology Response.** See response to comment 1.

**Peer Reviewer Dr. Ingersoll.** Appendix B. Insufficient analyses or summaries are presented in this appendix to review the data screening process (see Comment 1).

**Ecology Response.** Appendix B was revised and expanded. See response to comment 1.

**Peer Reviewer Dr. Ingersoll.** Page B-5. It is very surprising as to the types of chemicals that were screened out across toxicity endpoints (e.g., lots of metals of typical concern in sediment).

**Peer Reviewer Dr. Ingersoll.** Technical memorandum (dated March 14, 2010). The results of these analyses need to be described in the report. I do not remember discussing these analyses.

**Ecology Response.** This citation can’t be found anywhere in the report.

**Peer Reviewer Dr. Ingersoll.** Overview of the Biological Freshwater Sediment Standards (dated August 25, 2010): Hyalella azteca 10-d growth needs to be added as a sublethal endpoint. Biomass of amphipods and midge should also be added as endpoints.

**Ecology Response.** These endpoints could be required for reporting data and performance evaluated when enough data has been acquired, or site-by-site if helpful in evaluating bioassay results. In the RSET and SMS programs, changes to program endpoints can be made, but follow a specific process. An issue paper is prepared along with supporting data and other evidence, and presented at the public Sediment Management Annual Review Meeting. The federal and state sediment management agencies jointly consider the recommendation and make a decision after the meeting, considering public comment. Subsequent actions, such as rule revisions or database reprogramming, could take significantly longer. This approach would also be used for changes to the biomass endpoint (see below) or revision of the bioassay control criteria.

# Appendix D: Additional Comments and Responses

**Responses to Oregon DEQ Toxicology Workgroup Comments submitted in April 2010.**

The following are comments from the Oregon DEQ toxicology workgroup (Mike Poulsen, Jennifer Petersen) regarding Ecology's April 2010 proposed method of using the floating percentile method to develop sediment quality values (SQVs). Comments are shown in italics and responses in standard text.

**Comment. SL1 (low screening level)**
**Pro –** The approach of establishing an acceptable level of harm for bioassays, optimizing using all bioassay endpoints, and taking the lowest values for SL1 screening values may have merit as a method for developing potential SQVs.

**Ecology Response.** Thank you for your comments; we agree.

**Comment. Con –** Ecology's proposal to use higher levels (15% and 20% difference from control) to define harm for short-term mortality tests means that decisions will be made to leave sites with more than minor ecological harm. This is not consistent with the primary goal of low screening levels, which is to predict the absence of toxicity. Details regarding the development of SQVs using the floating percentile method have not been settled.

**Ecology Response.** The SMS rule was not designed to require proof of absence of toxicity. Rather the SMS were developed to manage environmental risk by minimizing adverse effects to a population or community of benthic organisms. WAC 173-204-315, "no adverse effects" for marine bioassay tests are defined using effects thresholds of 15-30% compared to reference, depending on the test endpoint. These thresholds are based on the minimum detectable difference in these bioassay endpoints. The freshwater SQVs were designed with the same principles in mind, except they are compared to the control, a more stringent standard. Therefore, 15-20% difference from control is well within the regulatory definition of "no adverse effects" for bioassay tests, and is based on minimum detectable difference, or the point at which an adverse effect can first be observed. Effects below this level are considered "no adverse effects," while effects exceeding this level are considered "minor adverse effects" up to the higher CLS/SL2. As a result, the primary goal stated above is met, as there are not expected to be biologically observable or statistically meaningful adverse effects below this level.

**Comment.** SL2 (high screening level). The goals for the level of harm associated with SL2 values are less clear than for SL1 values. It is not even clear if the level of harm at the SL2 level should be different from the level of harm at the SL1 level. PELs and PECs (i.e., SL2-type values) are based on the same level of harm (statistical difference) used for TELs and TECs (SL1-type values); the difference between PEL/PEC and TEL/TEC values is in the likelihood of harm. We can see that it may be an appropriate risk management decision to use a greater level of harm for SL2 values, given that the application of SL2 values is primarily for remediation decisions, and not investigation decisions associated with SL1 values.

**Ecology Response.** There are two equally valid approaches to establishing SQVs, both of which have been used by various state and federal programs. One is to establish two thresholds with differing probabilities of effects, and the other is to establish two thresholds with differing levels of effects. The approach used in the SMS rule is dictated by the narrative language of the rule, which defines two levels of effects – the “no adverse effects level” and the “minor adverse effects level.” Any numeric standards developed under the rule must reflect these two definitions; hence, the reason that this type of approach was selected.

**Comment.** Each SL2 value was developed by selecting a value other than the lowest value from all bioassay endpoints as long as it was greater than the lowest value. There appears to be no justification for this approach.

**Ecology Response.** The selection of the SL1 level is largely based on science – the minimum detectable difference as best it can be determined. The higher regulatory level is a risk management policy decision on the part of any agency based on best professional judgment by the experts – additional levels of risk or effects that can be tolerated when balanced against other issues. In this case, the entire distribution of possible values for each chemical fell within the statutory definition of “no adverse effects” to “minor adverse effects” and any one of them could have been selected. To determine the second-highest value Ecology followed the precedent set by the SMS marine standards, which have been in rule since 1991 and went through extensive scientific peer review, public review, and an EIS process. For the marine standards, the lowest AET was selected as the SQS/SL1 and the second-lowest AET was selected as the CSL/SL2, after consideration of all possible alternatives. This provides a practical range of values for site managers to work within, taking into consideration environmental protection, cost, and technical feasibility, while still ensuring that the final cleanup standard will not exceed “minor adverse effects.”

**Comment.** Existing Framework for Screening. We currently have sets of validated, national values to use (TELs/PELs and TECs/PECs). Any new set of screening values should be an improvement that meets the goals of the screening.

**Ecology Response.** When Ecology first embarked on development of freshwater SQVs, each of these existing SQV sets was evaluated to determine how well they predicted toxicity in Washington and Oregon sediments. The results showed very poor reliability in predicting the toxicity of freshwater sediments in the Pacific Northwest region (SAIC and Avocet 2002). This evaluation was repeated in 2008, with the same results (Avocet 2010). This evaluation is what prompted the agencies to consider development of an alternative approach to setting SQVs that would be more predictive. As was done for the SMS marine standards, Ecology ensured regional sediment samples were collected and analyzed using both Ecology and ASTM approved bioassays. Our goal all along has been to improve on the reliability of existing SQV sets. If at any time we had found that this was not the case, the development effort would have been discontinued in favor of the existing SQVs.

**Comment.** In the absence of defensible and validated screening values, we recommend using existing screening values. These values are currently used within a tiered sediment evaluation framework to evaluate sediment toxicity. This regulatory framework recognizes the appropriate application of sediment screening values within a larger evaluation framework that includes the interpretation of chemistry through the use of sediment quality screening numbers, empirical bioassay testing, and (ideally) changes in field communities to inform decision-making. Low SQVs are used to identify areas that need no further evaluation, and high SQVs are used to identify areas of probable toxicity. Areas that are above the low screen are not predicted to be toxic, but rather are not predicted to be non-toxic. The range between the low and high indicate areas that would benefit from further evaluation such as bioassay testing for evaluating toxicity.

**Ecology Response.** The sediment cleanup and dredging programs that will use these SQVs also apply them within a framework that includes optional biological testing and/or benthic evaluation. The framework described above is not actually our regulatory framework, but one proposed by a SETAC Pellston workshop and the developers of the TELs/PELs and TECs/PECs. The SMS rule framework that the current SQVs would be applied within differs somewhat in that it requires strong defensibility of the SQVs at the SL1 level, not only in predicting areas that are non-toxic, but in accurately predicting toxicity, so that both agency and regulated community resources are not wasted on unnecessary characterization efforts. The SL1 levels are used not only to screen sediments out, but to determine which sediments may be toxic and thus warrant further attention.

The TEL/TEC values have an 85-95% false positive rate, which indicates that minimal sampled stations would be screened out, eliminating the value of having any screening levels. While this is a conservative approach, Ecology must make risk management decisions based on reasonable predictions that sediment is contaminated in order to require further costly analysis and cleanup. Rather than use TEL/TEC type values, it would be more efficient to continue as Ecology has been, requiring bioassays at all stations for all projects. However, this approach is costly and does not provide predictability and consistency to make cleanup decisions. Alternatively, more accurate SQVs can and have been developed, and are proposed here.

**Comment.** The basic process should start with selection of biological endpoints that are considered relevant, with defined acceptable levels of harm. Proposed SQVs should be evaluated to see if goals are met. Modifying or dropping bioassay endpoints, revising definitions of acceptable harm, or otherwise altering evaluation methods in order to meet goals should not be done.

**Ecology Response.** The level of harm for the SL1 is defined as the minimum detectable difference**.** In the final 2010 proposal, we did not modify or drop bioassay endpoints, revise levels of harm, or otherwise alter evaluation methods. The question we asked late in the process was whether we had accurately determined what that minimal detectable difference actually was, once it became clear that our original interpretive guidelines gave results that seemed statistically difficult to interpret. The RSET workgroup agreed this might be important to pursue, but did not have the means to do so late in the process. Later, based on an internal peer review of Ecology's senior sediment experts and an independent recommendation by its external Sediment Workgroup in 2010 (SMS rule advisory group made up of state and nationally renowned sediment experts), Ecology undertook this evaluation.

The original proposal (10% difference from control) appeared to be within the statistical noise of the bioassay, and it was possible to test this by running a range of differences from control and observing the trends in statistical correlations and reliability results. We also checked with the members of the Sediment Workgroup, several local bioassay laboratories, and national experts, including Drs. Chris Ingersoll and Dave Mount. All of these people agreed that the original 10% difference was likely below the minimum detectable difference, and that 15-20% was probably more appropriate. These professional opinions were backed up by the results of the ANOVA analyses and reliability results, which showed improved statistical relationships between chemistry and bioassay results at slightly higher thresholds of difference. This conclusion appears reasonable considering that even the proposal for revised control mortality in these bioassays allows up to 15-20% mortality, and the variance among replicates in a test sample is likely higher than in a control sample. Under these conditions, a difference of only 10% from control is likely below the minimum detectable difference.

**Comment.** Ecology is proposing thresholds (15% or 20%) that allow greater short-term mortality of organisms, particularly at the SL1 level. Empirical bioassay results with short-term mortality between 10% and 20% would be considered adverse effects by DEQ toxicologists and EPA. This creates an inconsistency between the appropriate definition of harm and the thresholds used in the model to develop SQVs.

**Ecology Response.** Please see above discussion. We believe these thresholds represent the minimum detectable difference in our data set, and thus accurately define the boundary between no adverse effects and minor adverse effects, as required by the SMS rule. This is the point at which one begins to observe minor adverse effects, and appears consistent with the DEQ range (10-20%) cited above, as well as the existing marine SQVs.

**Comment.** As consistently applied in the DEQ Cleanup Program for the past ten years, a hit in any bioassay test would indicate harm for that sample location. Accordingly, reliability of the final set of SQVs should be evaluated using a pooled dataset. It is important to know the reliability of SQVs in predicting the results of all relevant bioassay endpoints.

**Ecology Response.** This is also true in the SMS rule at the SQS/SL1 level. Because the lowest of all the FPM values was selected as the overall SQS/SL1 value, it should be apparent that the combined SQS/SL1 level will be protective for all bioassays or combinations of bioassays that could be run at a site.

Running a pooled reliability analysis with a historical data set of this nature has a number of practical difficulties. Each station has a different combination of bioassays that were run, and some have only one or two. A pooled hit/no-hit determination for these stations is impossible, since in our current programs and the SMS, at least three bioassays are required. If three had been run at those stations, a different overall result might have been obtained. Use of these stations with limited bioassays to calculate FPM values for individual bioassay endpoints is possible and appropriate, but using them for a pooled analysis is not. However, leaving them out greatly reduces the size of the data set, and it is no longer of the same composition or representativeness as before. Moreover, there are 5 bioassays included in the complete data set, and it is not possible to know which of these might be used at any given site. Therefore, the results of a pooled analysis are not very meaningful. It is preferable to just use a conservative approach to selecting the final SQVs (the lowest of all the individual values), as has been proposed by Ecology, after determining that the values for each individual test and endpoint are reliable and sufficiently protective. This is fully consistent with the AET approach used to develop the SMS marine criteria.

**Comment.** The primary goal of low screening levels is to reliably predict the absence of toxicity. That is, the false negative rate must be acceptable. This means if we use the SQVs to justify no further evaluation at a site, we need to be reasonably confident there is a low error rate associated with this decision. Other goals (such as minimizing false positives) are secondary, and should only be considered once the primary management goal (false negative rate = 20%) is met. The false positive rate should not be improved at the expense of exceeding an acceptable false negative rate. Rather, the reason(s) behind that false positive rate should be investigated and resolved.

**Ecology Response.** The target and actual false negative rates for all bioassay endpoints fell within the targets set by the RSET workgroup and were not changed to achieve lower false positive rates or for other reasons. The model actually fixes the false negative rate and changes everything else in the optimization routines to ensure that the false negative target is met.

**Comment.** Reliability evaluations using the model dataset should not be used in place of validation. Any set of SQV values should be evaluated using a separate validation dataset, which we do not have. This is the only way to truly evaluate reliability without bias. Ideally, before adoption of SQVs, validation would also include correlating SQV values with changes in community structure (e.g. abundance and diversity) in the environment. This type of validation was done with nationally accepted and currently used values (Ingersoll 2001).

**Ecology Response.** Most SQVs are originally developed without validation, which occurs later once new data are available with which to conduct the evaluation. The reliability evaluation is not intended as a replacement for validation, but rather the best that can be done in the meantime. The reliability evaluation conducted for these SQVs is substantial and rigorous and has resulted in values more reliable than other SQVs. The RSET workgroup (including Oregon representatives) discussed whether to set aside a portion of the data set for validation or whether to include it all in the SQV calculations. It was decided to include all the available data in the SQV calculations, and to subsequently validate the SQVs once new regional data became available. Ecology still intends to carry out this process.

In the freshwater environments that exist in the Pacific Northwest, ranging from small mountain streams to large regional rivers, from dammed areas to channeled agricultural ditches, and from acidic lakes and wetlands to alkaline lakes and mining-impacted watersheds, it would be difficult to conduct validation against benthic community data. Most of these surveys do not have acceptable reference areas, and a 4-year RSET process to attempt to identify appropriate reference areas in the region largely failed. This type of evaluation (e.g., Ingersoll 2001) is really only possible in freshwater areas such as the Great Lakes, which are large and relatively uniform in their benthic environments and communities, more like Puget Sound.

**Comment.** We applied Ecology's SQVs to data from Portland Harbor, the largest sediment site in Oregon, using criteria acceptable to EPA and DEQ toxicologists (including 10% difference from control as acceptable difference, the correct method of calculating difference from control, and use of Hyalella biomass instead of growth results). Ecology's proposed screening values, both at SL1 and SL2 levels, did a poor job of predicting the absence of toxicity. If these SL1 values were used as screening levels, the conclusion would have been to walk away from one half the samples that were identified as toxic through empirical bioassay testing.

**Ecology Response.** A basic tenet of a defensible reliability assessment is that it must be conducted using the same biological endpoints used to calculate the SQVs; otherwise it is an apples and oranges comparison. The biological endpoints used to calculate the FPM values will also be used in Ecology's regulatory programs for consistency, and were reviewed and approved by the RSET biological testing subcommittee and Ecology's Sediment Workgroup. Portland Harbor is a highly complex, contentious, largely negotiated site where the methods and approaches used may or may not be representative of what would be used at other sites or projects. Ecology's SMS rule includes the ability to depart from the approaches described in the rule at any given site, of which Portland Harbor would be a good example. Partly because there is so much data with which to work. However, it cannot be assumed that these methods, thresholds, or other approaches will be applied to other sites in the region. An accurate reliability assessment of both the FPM values and the existing SQVs using bioassay interpretation guidelines consistent with those used to develop the SQVs is included in the SQV report.

**Comment.** SQVs developed using the FPM are developed as a set of values. However, evaluating the reliability of a set of numbers is distinctly different from evaluating the individual reliability of the chemical specific values. This is an important consideration, as SQVs will likely be applied on an individual chemical basis. The reliability of SQVs has not been evaluated on an individual chemical basis.

**Ecology Response.** The FPM SQVs are specifically designed to take into account covariance, additive/synergistic, and bioavailability issues by mimicking how chemicals are actually found in the environment, in mixtures. It will be clearly stated in our regulatory guidance that the SQVs are expected to be applied as a set (and not using quotients or other mathematical manipulations). It should be noted that most SQVs suffer from the opposite problem – they have been developed on an individual basis, and the simple mathematical methods used to combine them are not likely to accurately reflect the complex interactions between chemicals in the environment.

**Comment.** The prevalence of toxicity in a dataset (i.e., the ratio of hits to no-hits) will affect predicted-hit, predicted-no-hit, and overall reliability measurements. This makes it difficult to evaluate the meeting of goals for these measures.

**Ecology Response.** This is true, and is the reason that Ecology emphasizes false negatives and false positives in its decision-making, which are independent of the ratio of hits and no-hits in the data set. The predicted hit and predicted no-hit measures were largely included at DEQ's request for informational purposes. We believe overall reliability is still useful because the underlying data set is large and representative of the types of data likely to be generated in the future – including the general ratio of hits and no-hits or clean and contaminated sediments in the region. It is also one of the measures that is most easily understood by the public.

**Comment.** Inappropriate modification of parameters to meet goals. The acceptable level of harm is both a risk management decision and a technical decision that should be based on best available science. Once an appropriate level of acceptable of harm has been established, the results of developing any set of SQVs should be evaluated to see if the acceptable risk goal has been met. It is inappropriate to modify the acceptable risk level if the risk goal is not met. To do so would change the objective from meeting reliability limits for acceptable levels of harm in relevant bioassay endpoints to meeting reliability limits for higher levels of harm in some bioassays.

**Ecology Response.** To clarify, the acceptable risk level was not modified. The acceptable risk at the SQS/SL1 level is defined by the SMS rule in narrative terms as the "no adverse effects level." In technical terms, this has been interpreted and previously applied as the minimum detectable difference. Ecology's work to better define the minimum detectable difference has resulted in some relatively minor changes to the numeric bioassay interpretive guidelines, but these fall well within the ranges for other bioassays and are supported by both statistical evaluations and advisory opinions by local and national experts.

**Comment.** Definition of acceptable harm in bioassays. There is apparently substantial disagreement regarding the level of acceptable harm in bioassays. The written bioassay methods only discuss harm in terms of statistical significance from control or reference stations, and this approach (statistical significance only) was used to define harm in the development of TELs and TECs. Increasing toxicity thresholds, or definition of harm, can have implications for protection of benthic communities.

**Ecology Response.** The purpose of the thresholds at the SQS/SL1 level is not to allow additional or increasingly toxic levels of harm. Rather, they are intended to reflect the minimum detectable difference, consistent with the statistical significance approach described above. Programmatically, we could use statistical significance without a threshold. However, factors outside the applicant's control can sometimes affect the variability in the control sample (or the test sample), resulting in widely differing thresholds for different sites or even different bioassay batches within a site. For example, occasional control samples have been known to have no mortality in any replicate, making even the smallest degree of mortality a difference from control, while statistical significance in another batch might require a 30% difference. To make the decision process more predictable and fair, a consistent threshold representing a reasonable minimum detectable difference is used for all samples in all projects. This also provides better comparability between projects for the purposes of program and trend evaluation over time.

**Comment.** Adverse effects on bioassay test organisms, including mortality and sub-lethal effects such as growth impairment, have been correlated with community effects such as abundance colonization ability, emergence, and changes in community composition and diversity (EPA 2000, ASTM 2000). Therefore, adverse effects in a bioassay test to a single species are an indication of potentially significant adverse effects to a benthic community (EPA 2000).

**Ecology Response.** If by "significant," you mean "statistically significant," we would agree - hence, the use of the minimum detectable difference as the SQS/SL1 level.

**Comment.** We agree, however, that it may be acceptable to define harm at the SL1 level by a 10% difference from control (in addition to statistical significance). Ecology's proposal to use higher levels (15% and 20%) to define harm for short-term mortality tests means that decisions will be made to leave sites where ecological harm is considered moderate or severe by EPA definitions. DEQ toxicologists consider it unacceptable to base SL1 screening levels on a level of harm greater than 10%. We know of no other regulatory agency that has accepted these high levels of harm for low screening levels.

**Ecology Response.** Ecology would not consider a threshold of 15-20% moderate to severe effects. This would be inconsistent with the SMS rule, programmatic practice, and data suggesting that effects in this range represent the minimum detectable difference in most laboratory bioassays. As noted above, the marine biological SQS in Ecology's SMS rule are based on 15-30% thresholds of effects, representing minimum detectable differences. These levels are also used by the DMMP/RSET regional dredging programs. Washington State has evaluated these thresholds extensively over the years, conducting power analyses and other evaluations to ensure that the thresholds could be detected. This was not the case for all bioassays, suggesting that levels lower than these are not effectively implementable. It is likely that any program that is using statistical significance alone as its hit/no-hit criterion is achieving roughly these thresholds in practice.

While some SQVs use risk-based thresholds rather than effects thresholds due to statutory or mathematical differences from the SMS, thresholds higher than 10% are typically used in both cases. For example, British Columbia uses an $EC_{20}$ (20% effects level) for development of its sediment quality guidelines. EPA's Great Lakes National Program Office guidance recommended a difference from control of 20% (http://www.epa.gov/glnpo/arcs/EPA-905-B94-002, Chapter 6). Oregon's risk assessment guidance sets an Environmental Baseline Value at the $LC_{50}$ and requires that there be no more than a 10% chance that 20% of the population exceeds the $LC_{50}$ (which would require a multi-station evaluation). Benchmark values involving the TELs and TECs are often set at 0.2 (20% probability or incidence of a hit). While these approaches aren't entirely comparable to one another mathematically, we are not aware of any regulatory programs using a 10% or lower threshold on a programmatic (non-site-specific) basis.

**Comment.** The difference between comparisons with reference locations or laboratory controls should not be used as a justification for using higher levels of harm as acceptable levels. Reference locations in the Willamette River gave comparable if not better results than laboratory controls. Recent evaluations have shown little variability due to laboratories.

**Ecology Response.** Extensive evaluations of program data (both marine and freshwater) have shown that reference samples generally have equal or higher toxicity than control samples. On a programmatic basis, the average of reference samples is higher than control samples. In addition, reference samples (field samples) tend to have higher variability among replicates than control samples. Therefore, decreasing the statistical power of the comparison and increasing the minimum detectable difference. For these reasons, comparison to control is more conservative than comparison to reference, in general (though individual sites may vary). Because we are attempting to maintain program consistency between the SMS rule for marine and freshwater standards, the difference is worth noting. It would be inappropriate to have both a more conservative comparison framework and more conservative thresholds on one side compared to the other. This change makes them more consistent, though that was not its purpose.

**Comment.** In fact, laboratory results for conducting toxicity tests have shown that the current minimum control survival requirements for Hyalella and Chironomus are probably more lenient than necessary and may need to be raised (Ingersoll et al., 2008). As a result, test acceptability requirements for sediment toxicity tests currently outlined in ASTM (2000) and USEPA (2000) will likely be revised in the next iteration. This work shows that variability in results is likely more a function of environmental bioavailability rather than laboratory “noise”.

**Ecology Response.** We are aware of the potential for this upcoming modification, and have factored it into our deliberations. If the control mortality limits are reduced by 5-10%, then the overall mortality that could be allowed in the test sample (allowable control mortality + allowable difference from control) would remain the same even if the thresholds are adjusted to more accurately reflect the minimum detectable difference between the sample and the control.

**Comment.** Variability (“noise”) described in the data can be an important reflection of real variability in bioavailability in different sediment environments with similar chemical concentrations. The outcome of moving the toxicity threshold to a higher percent difference from control does not address this problem, and only effectively changes “hits” to “no-hits” without evaluating the probability that those samples are actually now false negatives within the database. If the resulting SQVs are used for screening, sites falling below new higher threshold SL1 levels will not be evaluated further using empirical bioassay testing to determine toxicity. Bioassays are a stronger line of evidence for toxicity than SQVs.

**Ecology Response.** Because the new thresholds more accurately reflect the actual minimum detectable difference in these bioassays, the interpretive outcomes would likely be similar either way. Indeed, as noted by one DEQ reviewer, the proposed chemical SQV values resulting from this change have actually become more conservative, not less. This reflects the basic principle that reducing error in the overall model results in more accurate SQVs without sacrificing protectiveness.

**Comment. Use of pooled data to evaluate reliability.**
The reliability of screening values should be evaluated using the entire set of bioassays (called pooled data). There has been lack of agreement on this key issue. In making a decision at a DEQ cleanup site, toxicity in any bioassay performed on a sediment sample would indicate harm. We therefore want to know the reliability of SQVs in predicting the results of any relevant bioassay endpoints. This approach is consistent with that used by MacDonald et al. (2000) to evaluate the accuracy of their TEC and PEC predictions. They considered a hit in any bioassay a hit for the sample. In addition to Hyalella and Chironomus, they considered Hexagenia, Lumbriculus, Ceriodaphnia, and Microtox.

EPA toxicologists, EPA project managers, their expert consultants including Don MacDonald and Jay Field, and the Lower Willamette Group all agree with DEQ toxicologists that a pooled set should be used to evaluate reliability.

**Ecology Response.** Please see the response to Comment 3 above for why a pooled analysis was not conducted and is not recommended. By selecting the lowest of the individually reliable FPM values as the SQS/SL1, Ecology believes that the resulting values will be at least as sensitive as any of the individual values in detecting toxicity.

**Comment.** Primary goal of low screening levels. The primary objective of SL1 screening values is to predict the absence of toxicity. This corresponds to having an acceptable low false negative rate and/or false predicted no-hit rate. Anything else is secondary. This objective is consistent with national screening levels, such as TECs developed by MacDonald and Ingersoll. In his September 2008 review of the predictive models for the Portland Harbor site, MacDonald states that low screening values should be considered reliable if there is a low incidence of toxicity (i.e., <10%) for sediment samples that have chemical concentrations below the screening levels for all measured substances. MacDonald states that the low screening levels should not necessarily be evaluated to determine how well they predict toxicity because SL1 levels are designed to predict the absence of toxicity.

**Ecology Response.** Ecology differs with this statement describing the purpose of the SL1 values, and further notes that a screening level is only effective if it does screen out a reasonable percentage of samples. Excessively high false positive rates do not achieve that goal. The MacDonald approach to shield the TEL/TEC values from evaluation of both types of error may reflect the poor performance of these values in this regard.

**Comment. Validation of SQVs.** Our recommendation is to use an appropriate validation dataset to evaluate reliability. This was a specific recommendation by Don MacDonald in his 2008 review of the floating percentile and logistic regression methods of developing SQVs for Portland Harbor. He suggested a validation approach where we optimize on ¾ of the data, and use the remaining ¼ of the data to test the accuracy of the predictions. We suggest that Ecology consider the importance of conducting a validation evaluation prior to establishing SQVs in rule.

**Ecology Response.** Please see the discussion of this point under Comment 5 above. Validation is a valuable exercise, but not at the cost of ¼ of the existing data. This will be conducted in the future as soon as sufficient data are available. The RSET workgroup carefully considered this point, and decided to proceed with adoption, followed by validation studies as agency resources allow.

**Comment.** Application of proposed SQVs to Portland Harbor data. We evaluated the reliability of the SQVs using the Portland Harbor dataset. This is a simplistic evaluation that is short of a validation. The Portland Harbor data were used in the derivation of Ecology's SQVs, so it is not an independent dataset. Plus, we used EPA's definitions of harm, which included a 10% difference from control, calculated using the correct method as specified by EPA. This is not Ecology's definition of harm. But it was a convenient evaluation to run, and somewhat relevant (because it applied proposed regional levels to data from a large Oregon site). Here are the results, shown with national screening values for comparison.

| | % False | % False | % Predicted | % Predicted |
|---|---|---|---|---|
| SL1 | Negative | Positive | Hit | No Hit |
| Ecology SL1 | 50 | 46 | 39 | 65 |
| TEC | 8 | 81 | 40 | 80 |
| SL2 | | | | |
| Ecology SL2 | 60 | 8 | 61 | 84 |
| PEC | 49 | 25 | 37 | 84 |

The primary goal for SL1 is a false negative rate less than 20 percent. Using EPA's definition of harm, Ecology's SL1 SQVs result in 50 percent false negatives. Of the known toxic areas, the proposed screening values will miss half of them. For SL2, we can focus more on reducing the false positive rate, but we still want FN = 20%. Ecology's SL2 SQVs still result in a highly unacceptable false negative rate (FN = 60%). EPA's definition of harm at the SL2 level (25% difference from controls) is mostly the same as Ecology's definition.

**Ecology Response.** This is not an appropriate reliability evaluation, as the biological interpretive endpoints being used are much different and more conservative than those on which the chemical SQVs were based. This approach would result in a perception of high false negatives, since it would not be possible for chemical SQVs to predict levels of effects for which they were not designed. However, the levels of false negatives are "locked in" as part of the model and are not allowed to vary. They do not exceed the target levels established by the RSET workgroup, providing further evidence that the above reliability analysis is not accurate. However, it is worth noting the extremely high false positive rate of the TECs, which accords with our own evaluation.

An accurate reliability analysis comparing the TECs/PECs and other existing guidelines to the proposed FPM values is provided in the SQV Report, using consistent definitions of biological effects and chemical SQVs. Both the SQVs and the biological effects interpretive guidelines would be promulgated at the same time, to ensure consistency in the regulatory approach.

**Comment.** Individual chemical reliability. There is an implied assumption if reliability is met by a combination of SQV values that the reliability of the individual chemicals is also acceptable. The floating percentile method is designed to optimize screening values for a set of chemicals; the FPM is not designed to optimize reliability of screening values for individual chemicals, and this point has been acknowledged throughout the FPM development process. Nevertheless, we are concerned that SQVs will likely be applied on an individual chemical basis. At many sites, the primary risk drivers are one or two chemicals. Our initial evaluations show that individual chemical SQVs developed using the FPM have unacceptably high false negative rates. TECs are developed on an individual chemical basis, and have better false negative rates.

**Ecology Response.** Please see the response to Comment 7 above. There is no such implied assumption, and the implementation guidance will be clearly written to identify this point.

However, this does point out a key difference between the two approaches. It is because the individual TECs are designed to each have low false negative rates (although the biological hits for each may in fact be due to other chemicals) that the TECs as a whole have such high false positive rates.

**Comment.** Influence of toxicity prevalence on reliability measures. Given the dataset used to derive both SL1 and SL2 screening values, it is highly unlikely to meet the false negative and false positive goals of 20%, and also reach the predicted hit reliability goal of 80% for SL2 values.

**Ecology Response.** For nearly all bioassay endpoints at both effects levels, the above goals were met, as reported in the SQV report. All of the bioassay endpoints met the reliability target ranges set by the RSET workgroup.

**Comment.** This is because the dataset contains mostly no-hit results. Even with a good false positive rate, because there are many no-hit samples, there will be many predicted-no-hit samples relative to the number of predicted-hit samples. Teresa pointed out early in the process that predicted-hit and predicted-no-hit measures depend on the mix of hit and no-hit results.

**Ecology Response.** This is true. However, false positive and false negative rates are independent of the percentage of hits and no-hits in the database, and therefore this ratio does not affect the primary goals above substantially.

**Comment.** We also considered another issue that will affect reliability measurements. In using a high screen to identify areas requiring remediation, the samples are more likely to be contaminated than would be expected from the general database. The actual predicted-hit reliability rate, when calculated using a more contaminated dataset, should be greater than the predicted-hit reliability calculated using the SQV spreadsheet.

**Ecology Response.** As discussed above, Ecology does not consider the SQS/SL1 biological thresholds a high screen, which is consistent with the SMS rule. It is interesting to note that the predicted hit values may vary under different circumstances, but these are not the primary reliability measures used for decision-making.

**Comment.** In fact, false negative and false positive rates could also change when a more contaminated dataset is used. The false negative rate as calculated in the SQV spreadsheet could be overestimated at the SL2 level (and underestimated at the SL1 level).

**Ecology Response.** It is not clear how false negative or false positive rates could change with a more contaminated data set. This data set spans more than 10 years of cross-program history and we believe it is representative of the types of data that will be collected in the future and the levels and percentages of contaminated sediments that will be encountered. Sediment data are collected and SQVs are used for a wide variety of purposes, including cleanup sites (before and after), dredging projects, ambient monitoring, urban bay assessments, NPDES permits, disposal site monitoring, lease transfers, etc. Therefore, a significant change in the nature of the underlying data set does not seem likely. However, the overall concern lends support to using a regional data set rather than a national one in calculating SQVs, as different areas of the country may differ in these respects.

**Comment.** Covariance. There is known variability of SQV results given that the floating percentile method is a multivariate model with more variables than constraints. As explained, there are multiple solutions that provide the same reliability, especially for chemicals that are strongly correlated with each other. If two chemicals co-vary, the model may set the screening level for the first chemical low and the level for the second chemical high, or the level for first chemical high and the level for the second chemical low, and calculate the same reliability. By combining results from all the relevant bioassay endpoints, we may be able to see the range of suitable screening values for the chemicals. It is possible that the lowest (or highest) concentrations are artifacts of the approach used to derive the screening values. It would take considerable effort to demonstrate that the low screening concentrations resulting from individual bioassays are artifacts. However, unless this is shown, we do not consider it appropriate to dismiss the relevance of the low screening values.

**Ecology Response.** Covariance is an issue that affects all data sets and SQV calculation efforts. Most methods do not take it into account at all, and one helpful aspect of the FPM is the ability to see it in the results, even if there is no perfect way of dealing with it. As suggested in the comment, in Ecology's current proposal, all of the FPM values are retained and the lowest ones are used to set the SQVs.

**Comment.** Bioavailability. We suggest that Ecology look at sources of variability, such as bioavailability. Jay Field of NOAA found that organic carbon normalization is important for improving relationships using the logistic regression method for developing sediment SQVs. It has been shown that similar concentrations of a chemical in units of mass of chemical per mass of sediment dry weight often exhibit a range in toxicity in different sediments (DiToro et al., 1991; USEPA, 1992).

**Ecology Response.** This has been done in the past and in the current calculation effort. TOC-normalization was tested side-by-side with dry weight values in 2003 and in 2008, as well as for a number of other projects, listed in the SQV report. These evaluations did not improve the reliability of the results. The marine AETs are the only state or provincial values we are aware of that are TOC-normalized, and it was recognized at the time they were developed that the dry weight AETs were equally predictive. TOC-normalization of the marine AETs was largely done in deference to theory, but it has not been demonstrated to improve predictiveness. In addition, there have been significant implementation difficulties associated with the OC-normalized values, including sediments that had TOC outside the equilibrium partitioning range (too low or too high), sediments with excessive TOC of anthropogenic origin, and general difficulties explaining the approach to the regulated public.

A number of other summing and normalization approaches were also tested, with the result that some chemical classes were summed. Other approaches did not improve reliability.

Bioavailability will always differ between sediment samples due to the origin and nature of the chemicals themselves and the many different physical and chemical aspects of the sediment and porewater. A considerable effort has been focused on better understanding the contribution of some of these different factors, such as OC. Ultimately, if there were clear evidence that these other ways of looking at the chemistry served to consistently improve the predictive accuracy of the SQGs, then we would adopt a different approach. The FPM is the only method for developing SQGs that explicitly deals with the bioavailability of a chemical in a synoptic data set by managing its effect on false negatives and false positives. Also, our regulatory framework with a range between the lower and upper regulatory levels and a biological override allows appropriate consideration of site-specific factors where there is reason to believe the SQGs may be less effective predictors.

**Comment.** Bioassay Endpoint Normalization. Endpoint normalization to control was found to be an important consideration, at least for the Portland Harbor database. The procedure used for normalizing to control is important in reducing variability in replicate response that can confound the calculation of statistical difference. Instead of RSET's current approach of subtracting the test result (T) by the control response (C), EPA and other agencies routinely normalize control response by dividing the (T) by (C). For example see (http://www.epa.gov/glnpo/arcs/EPA-905-B94-002/B94002-ch9.html#RTFToC83).

**Ecology Response.** To clarify, the cited report does not support this recommendation. In Chapter 6, the data analysis approach for bioassays describes a difference from control of 20% as being indicative of adverse effects. We have had this discussion over the years in the DMMP and SMS programs and have settled on a difference from control as the best way of representing the minimum detectable difference concept we are trying to represent. We are not aware of any data supporting one approach as better than another, or of programs that divide mortality results by the control response. Because the difference method was consistent with the SMS rule and approved by the RSET workgroup, this approach was programmed into our EIM database at the time it was developed. As a result, it is hard-coded to work with mortality data, which cannot be divided by the control (it would have to be in units of survival). Reprogramming would be a significant effort that would only be undertaken if it could be clearly justified, as it affects all of the bioassay hit/no-hit and statistical comparison routines.

As noted above, approaches taken at an individual site, especially a large, complex Superfund site, may be different from those used on a routine basis. Portland Harbor has a particularly difficult data set, and it is not surprising that alternative interpretive approaches were needed. It is not clear that this modification would be useful in other cases; a more thorough evaluation would be needed involving more data and other sites.

In the RSET and SMS programs, changes to program endpoints can be made, but follow a specific process. An issue paper is prepared along with supporting data and other evidence, and presented at the public Sediment Management Annual Review Meeting. The federal and state sediment management agencies jointly consider the recommendation and make a decision after the meeting, considering public comment. Subsequent actions, such as rule revisions or database reprogramming, could take significantly longer. This approach would also be used for changes to the biomass endpoint (see below) or revision of the bioassay control criteria.

**Comment.** Biomass as a bioassay endpoint. The use of biomass as a bioassay endpoint was also found to be important for the Portland Harbor database. Biomass is presented instead of growth in order to use reduce confounding factors to growth by mortality in bioassay tests (Call et al., 1999). Experts acknowledge that because the growth endpoint is not independent of survival, looking at growth by itself can be misleading. For example, if organisms die during the test, there is more food available for the other surviving organisms. These organisms can grow larger than they otherwise would. To reduce these confounding factors, a biomass endpoint is calculated as the total dry weight of surviving organisms (the RSET methodology has growth as per individual organism) in each replicate exposure chamber. A number of experts (e.g., Dave Mount, EPA; Chris Ingersoll, USGS; Don MacDonald, MESL) currently recommend using the biomass endpoint (total mass of survivors in test sample vs. control) as a way to combine growth and survival endpoints for the Hyalella 28-day and Chironomus 10-day tests.

**Ecology Response.** We are aware that there is a movement toward using biomass as opposed to average growth per individual. However, it has not yet been adopted by ASTM or regionally through the process described above, and control limits are not yet available for this endpoint. Ecology will consider revising the growth endpoints once these programmatic steps have been taken.

In terms of the SQVs, the problem was resolved in this data set by removing growth data that may have been unduly influenced by mortality from the data set. Therefore, the resulting SQVs should be usable with either endpoint, particularly as there is no difference between them for most samples.

**References**

Avocet. 2011 (SQV report). Development of Benthic SQVs for Washington, Oregon, and Idaho. Prepared by Avocet Consulting, Olympia, WA for the Washington Department of Ecology and Oregon Department of Environmental Quality on behalf of the Regional Sediment Evaluation Team.

SAIC and Avocet. 2002. Development of Freshwater Sediment Quality Values in Washington State, Phase I Final Report. Prepared by SAIC, Bothell, WA and Avocet Consulting, Kenmore, WA for the Washington Department of Ecology, Olympia, WA.

SAIC and Avocet. 2003. Development of Freshwater Sediment Quality Values for use in Washington State. Phase II Final Report: Development and Recommendation of SQVs for Freshwater Sediment in Washington State, Sept. 2003, Publication No. 03-09-088, Prepared by SAIC, Bothell, WA and Avocet Consulting, Kenmore, WA for the Washington Department of Ecology, Olympia, WA.

**For questions contact: Russ McMillan, WA Department of Ecology, (360) 407-7536, Russ.McMillan@ecy.wa.gov**